

SERIES TEMPORALES

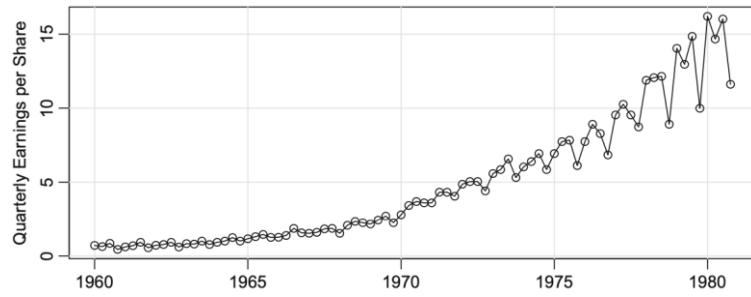
Iker Caballero Bragagnini

Tabla de contenido

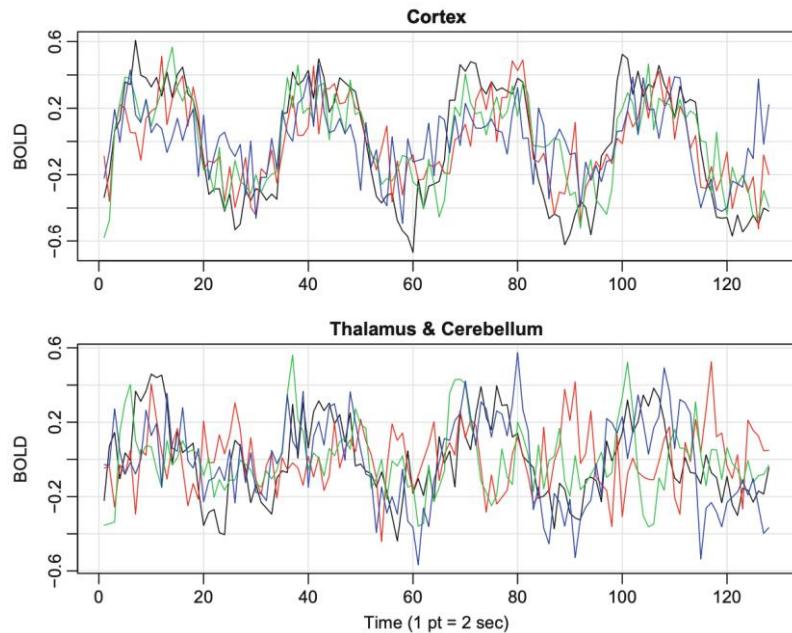
LAS CARACTERÍSTICAS DE LAS SERIES TEMPORALES	2
LAS REGRESIONES CON SERIES TEMPORALES Y ANÁLISIS EXPLORATORIO	26
LOS MODELOS ARIMA: LOS MODELOS ESTACIONARIOS	38
LOS MODELOS ARIMA: LA AUTOCORRELACIÓN	49
LOS MODELOS ARIMA: LA ESTIMACIÓN	56
LOS MODELOS ARIMA: LA PREDICCIÓN	59
LOS MODELOS ARIMA: LOS MODELOS NO ESTACIONARIOS	67
LOS MODELOS ARIMA: LA IDENTIFICACIÓN Y EL DIAGNÓSTICO	72
LOS MODELOS ARIMAX: ANÁLISIS DE INTERVENCIÓN Y DE VALORES ATÍPICOS	85
LOS ELEMENTOS BÁSICOS DE LA TEORÍA DEL DOMINIO TEMPORAL	100
LOS MODELOS DE ESTADO-ESPACIO: EL MODELO LINEAL NORMAL	105
LOS MODELOS DE ESTADO-ESPACIO: KALMAN Y ESTIMACIÓN	110
LOS MODELOS DE ESTADO-ESPACIO: DATOS PERDIDOS Y MODELOS ESTRUCTURALES	121
LOS MODELOS DE ESTADO-ESPACIO: ERRORES CORRELACIONADOS Y <i>BOOTSTRAP</i>	127
LOS MODELOS DE ESTADO-ESPACIO: MODELOS LINEALES DINÁMICOS Y VOLATILIDAD ESTOCÁSTICA	130

Las características de las series temporales

- El análisis de datos experimentales que se han observado en diferentes puntos en el tiempo deriva a problemas nuevos y únicos en el modelaje estadístico y en la inferencia
 - La correlación obvia introducida por el muestreo de puntos adyacentes en el tiempo puede restringir severamente la aplicabilidad de métodos estadísticos convencionales que tradicionalmente dependientes de la suposición de independencia y distribución idéntica (iid)
 - El enfoque sistemático por el que uno pasa de responder las preguntas estadísticas y matemáticas que generan las correlaciones temporales se denomina análisis de series temporales
 - La motivación del análisis de series temporales es la descripción y la predicción de series temporales, en donde hay cuatro objetivos principales: la descripción, la estimación, la validación y la predicción
 - La descripción consiste en describir patrones temporales en las series temporales, tales como efectos estacionales, la ciclicidad, las tendencias, los valores atípicos, los *breaks*, etc.
 - La estimación, que consiste en estimar los valores de los parámetros de las series temporales
 - La validación, que consiste en validar los parámetros estimados y decidir si estos son significativos o no
 - La predicción, que consiste en predecir valores futuros para las series temporales
 - El impacto del análisis de series temporales en aplicaciones científicas se puede ver a través de mencionar campos en donde surgen los problemas de series temporales
 - Muchas series temporales ocurren en el campo de la economía y las finanzas, en donde se lida con figuras mensuales de desempleo o con cotizaciones de bolsa diarias

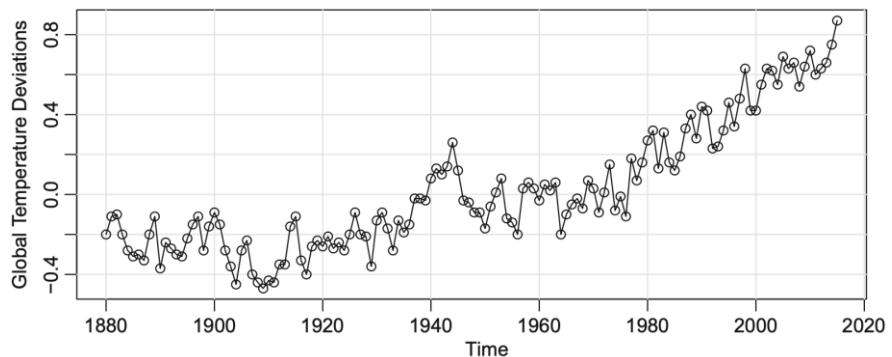


- Los científicos sociales lidian con series temporales de la población, tales como tasas de natalidad o de mortalidad
- En ciencias naturales también se tiene interés en varias series temporales, tales como medicina, epidemiología, etc.



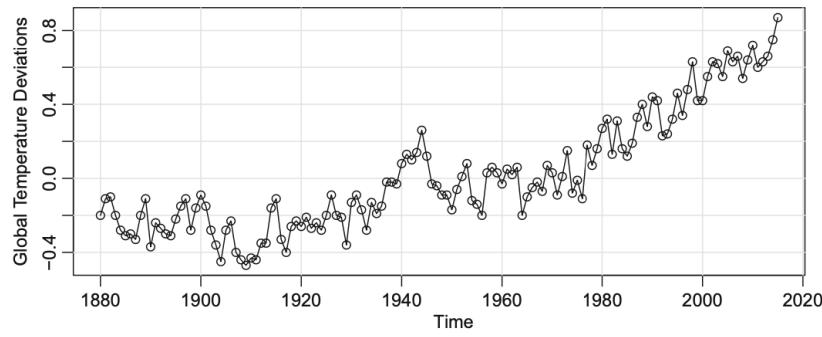
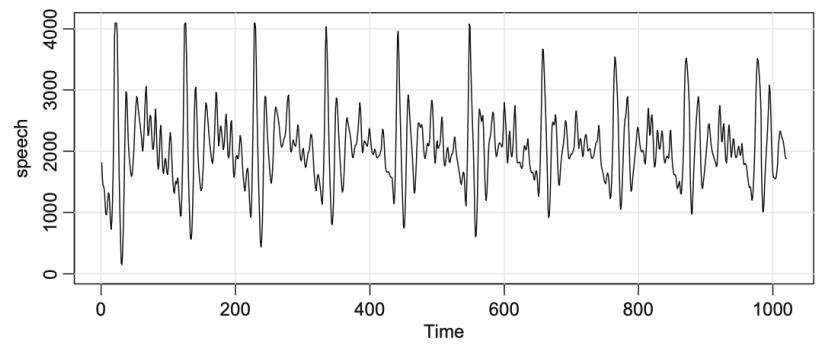
- El primer paso en cualquier tipo de series temporales siempre involucra la examinación minuciosa de los datos recogidos en el tiempo
 - Este escrutinio suele sugerir el método de análisis y los estadísticos que se usarán para poder resumir y caracterizar la serie temporal
 - Existen dos enfoques principales para el análisis de series temporales: el enfoque del dominio temporal, que se enfoca en investigar las relaciones de los retrasos o *lags*, y el enfoque del dominio de frecuencia, que se enfoca en investigar los ciclos
- El objetivo principal del análisis, por tanto, es desarrollar modelos matemáticos que proporcionen descripciones plausibles de los datos, por lo que es necesario establecer un marco matemático inicial con el que trabajar

- Con tal de proporcionar un marco estadístico para describir el carácter de los datos que parecen que fluctúan de manera aleatoria en el tiempo, se asume que las series temporales se pueden definir como una colección de variables aleatorias indexadas acorde al orden en el que los datos se han obtenido
 - En general, una colección de variables aleatorias $\{x_t\}$ indexada por t se conoce como un proceso estocástico. El índice t normalmente será discreta y varía sobre los enteros $t = 0, \pm 1, \pm 2, \dots$ o un subconjunto de los números enteros
 - Los valores observados de un proceso estocástico se conocen como realizaciones de los procesos estocásticos
- Es convencional representar la serie temporal de manera gráfica a través de poner los valores de las variables aleatorias en el eje vertical y la escala temporal en el eje horizontal



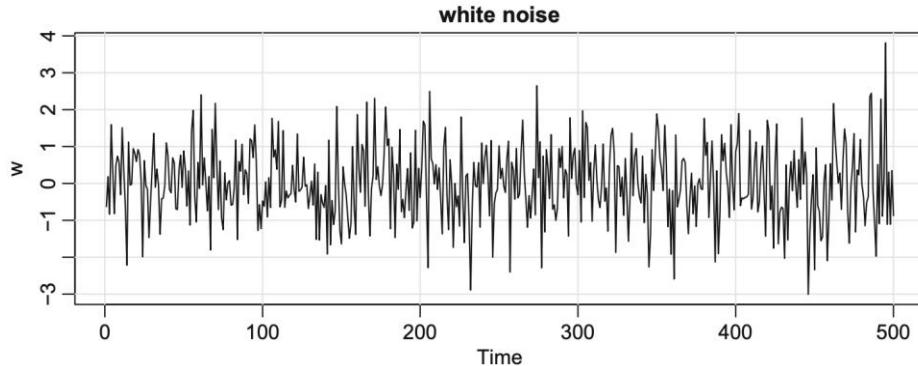
- Además, es usualmente conveniente conectar los valores en periodos adyacentes con tal de reconstruir visualmente la serie temporal continua hipotética que puede haber producido esta muestra discreta
- Muchas series temporales podrían observarse en cualquier punto continuo del tiempo, por lo que se consideran series temporales continuas
- La aproximación de estas series temporales se conoce como series de parámetros temporales discretos, de modo que se muestrea en puntos temporales con un espacio constante
 - Esta aproximación reconoce el hecho de que la mayoría de observaciones son discretas por las restricciones inherentes en la recolecta de datos

- No obstante, esto no quiere decir que el intervalo o la tasa de muestreo (como se recogen las observaciones) no es importante, dado que la apariencia de los datos puede cambiar completamente si se adopta una tasa de muestreo insuficiente. Esta distorsión en la apariencia se denomina efecto de alias o *aliasing*
- Las características visuales fundamentales entre diferentes series temporales son los diferentes grados de suavidad o *smoothness* que se pueden ver



- Una posible explicación para esta suavidad es que es inducida por la suposición de correlación entre puntos adyacentes en el tiempo, de modo que x_t puede depender de valores pasados x_{t-1}, x_{t-2}, \dots
- Este modelo expresa una manera fundamental en la que se puede pensar en generar series temporales realistas, de modo que, para empezar a desarrollar colecciones de variables aleatorias para modelar series temporales, se pueden considerar modelos fáciles como el de ruido blanco, la media móvil, la autorregresión o un camino aleatorio
- Algunos de los modelos matemáticos más utilizados y en los que se profundizará son los siguientes:
 - Un tipo de serie temporal puede ser una colección de variables aleatorias w_t no correlacionadas con media nula y con varianza

constante σ_w^2 , y el proceso estocástico resultante ruido blanco, de modo que se denota este proceso aleatorio se denota como $w_t \sim wn(0, \sigma_w^2)$

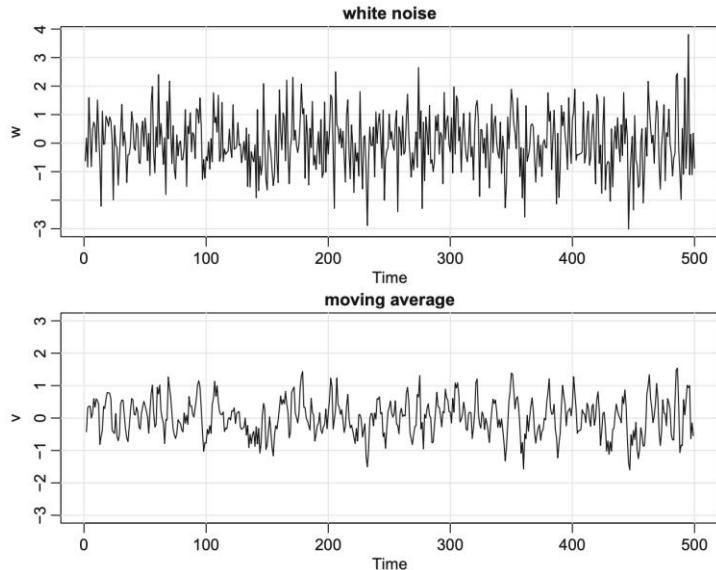


- Las series temporales generadas de variables no correlacionadas se utilizan mucho en ingeniería, y el uso del adjetivo “blanco” viene dado por la analogía con la luz blanca, e indica que todas las oscilaciones periódicas posibles están presentes con la misma fuerza
- A veces se requerirá que el ruido sean variables aleatorias independientes e idénticamente distribuidas con media 0 y varianza σ_w^2 , denotándolo por $w_t \sim iid(0, \sigma_w^2)$. Un caso particular muy útil es el ruido blanco gaussiano, en donde $w_t \sim iid N(0, \sigma_w^2)$
- Si el comportamiento estocástico del ruido blanco pudiera explicar la mayoría de series temporales, los métodos estadísticos tradicionales bastarían
- Es posible reemplazar w_t por una media móvil que suavice la serie. Por ejemplo, se puede calcular la media aritmética de w_t junto a observaciones vecinas tanto en el pasado como en el futuro, lo cual lleva a obtener el siguiente modelo:

$$v_k = \frac{1}{k} \sum_{j=-k}^k w_{t+j} \quad \text{for } t = 1, 2, \dots, n$$

- En este caso, k denota el número de observaciones que se utilizan en la media móvil y n es el número total de observaciones
- Como se puede ver, la serie temporal producida por la media móvil es una versión suavizada de la serie anterior, reflejando el hecho de que las oscilaciones más lentas son más aparentes y las oscilaciones más rápidas se eliminan (el efecto de hacer la media

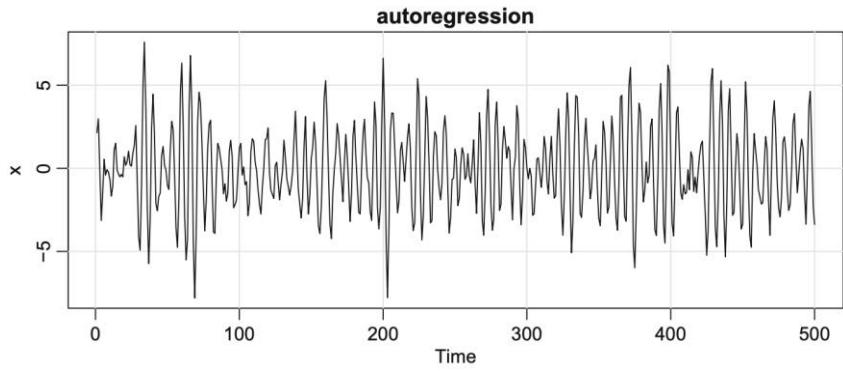
con este tipo de observaciones, eliminando el efecto de observaciones muy grandes o muy pequeñas)



- Una combinación lineal de valores en una serie temporal tal como la de la media móvil, genéricamente, se conoce como una serie filtrada
- También se puede utilizar una ecuación que represente una regresión para el valor actual x_t en términos de sus valores pasados y que incluya un componente estocástico como w_t . Este modelo descrito se conoce como autorregresión

$$x_t = \phi_0 + \sum_{i=1}^k \phi_i x_{t-i} + w_t \quad \text{for } t = 1, 2, \dots$$

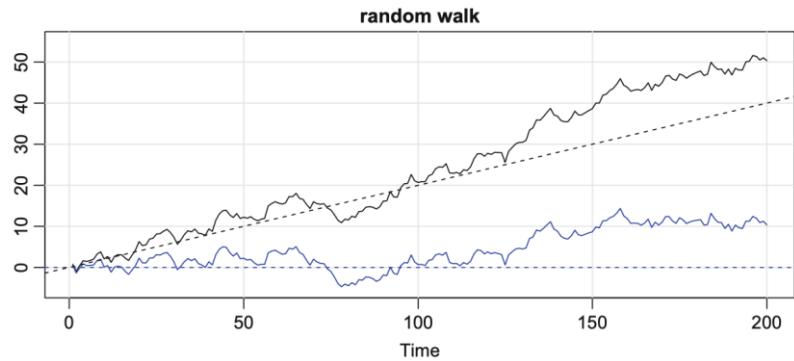
- En este caso, k es el número de valores pasados o retrasos de x_t en la autorregresión
- Un problema con este modelo es que varía mucho dependiendo de las condiciones iniciales, pero teniendo unos valores iniciales, es posible obtener series de la variable de resultado sustituyendo los valores predichos en los retrasos



- Este tipo de modelo y sus generalizaciones se utilizan mucho como modelos subyacentes para muchas series
- Un modelo para analizar la tendencia que se observa en una serie temporal es el modelo de camino aleatorio con tasa de deriva o *random walk with drift*, el cual tiene la siguiente forma:

$$x_t = \phi_0 + x_{t-1} + w_t \quad \text{for } t = 1, 2, \dots$$

- Este modelo se define para todos los periodos $t = 1, 2, \dots$ y se asume que hay una condición inicial x_0 , donde w_t es ruido blanco. La constante ϕ_0 se conoce como tasa de deriva o *drift*, y cuando $\phi_0 = 0$, este modelo se conoce como camino aleatorio
- El nombre de camino aleatorio proviene del hecho de que, cuando $\phi_0 = 0$, el valor de la serie temporal en el momento t es el valor de la serie temporal en el momento $t - 1$ sumado a un movimiento completamente aleatorio determinado por w_t . Cuando $\phi_0 \neq 0$, entonces las observaciones siempre sufren un aumento constante de ϕ_0



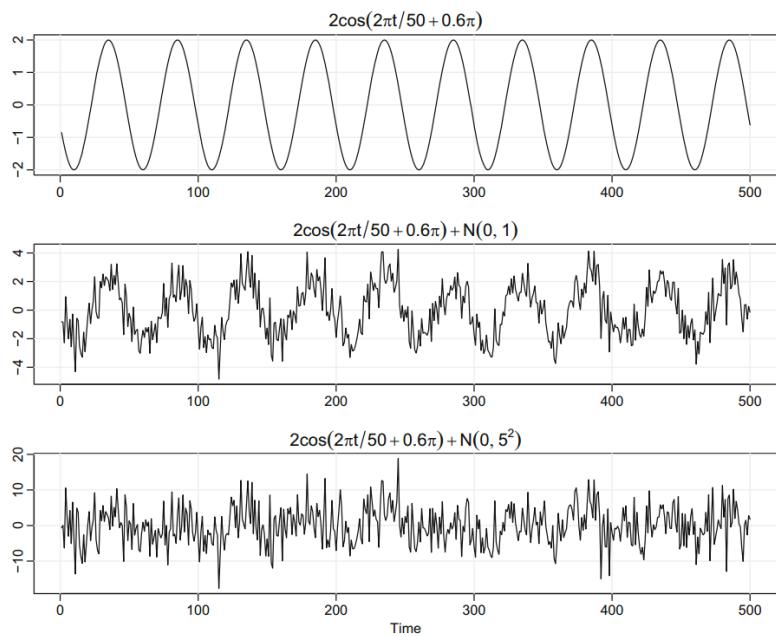
- Es posible reescribir el modelo como una suma acumulada de observaciones de ruido blanco para $t = 1, 2, \dots$

$$x_t = \phi_0 t + \sum_{j=1}^t w_j \quad \text{for } t = 1, 2, \dots$$

- Muchos modelos realistas para generar series temporales asumen una señal subyacente con una variación periódica constante, contaminada por la adición de un ruido aleatorio
 - De este modo, se puede considerar un modelo matemático para la serie temporal x_t que depende de un término de ruido w_t pero no de valores futuros ni pasados. Un ejemplo es la siguiente función:

$$x_t = A \cos(2\pi\phi) + w_t$$

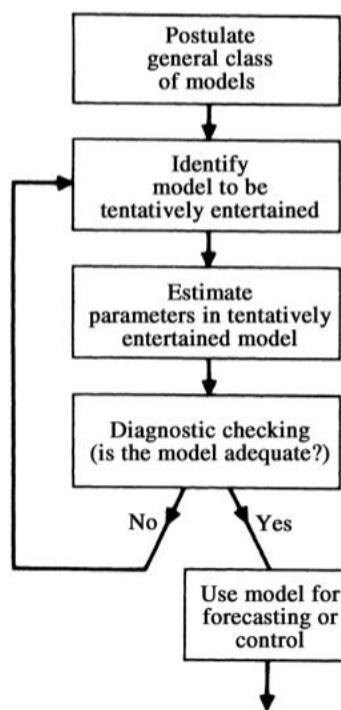
- Un valor más grande de σ_w hace que la señal se “oscurezca” más, lo cual quiere decir que hay más ruido y que por tanto se distingue menos



- El cuánto se oscurece la señal depende de la amplitud de la señal y del tamaño de la σ_w . La *ratio* de la amplitud de la señal con respecto a σ_w (o cualquier función de la *ratio*) se suele llamar *signal-to-noise ratio* (SNR), y cuanto mayor es la *ratio*, más fácil es detectar la señal
- Más adelante se podrá utilizar el análisis espectral como una posible técnica para detectar señales regulares o periódicas. En particular, se enfatiza la importancia de modelos aditivos tales como $x_t = s_t + v_t$ en donde s_t es la señal y v_t denota una serie temporal que puede ser un proceso de ruido blanco o una serie correlacionada

- Las ideas básicas del modelaje de series temporales son la parsimonia y los niveles iterativos en la selección del modelo

- En la práctica, es importante emplear el menor número posible de parámetros para unas representaciones adecuadas, lo cual se denomina principio de parsimonia. El uso de muchos parámetros deriva en estimaciones pobres
- Para la creación de modelos, se suele utilizar conocimiento teórico incompleto para indicar la clase de funciones matemáticas a utilizar y se ajustar empíricamente a los datos estas funciones. En general, se suele aplicar un enfoque iterativo para la construcción de un modelo:



- Una descripción completa de las series temporales, observada como una colección de n variables aleatorias en puntos arbitrarios t_1, t_2, \dots, t_n para cualquier entero positivo n , se proporciona por la función de distribución conjunta

$$F_{t_1, t_2, \dots, t_n}(c_1, c_2, \dots, c_n) = P(x_{t_1} \leq c_1, x_{t_2} \leq c_2, \dots, x_{t_n} \leq c_n)$$

- Desafortunadamente, las funciones de distribución multivariante no pueden escribirse fácilmente a menos que las variables sean conjuntamente normales
 - Aunque la distribución conjunta describe completamente los datos, es una herramienta poco útil para representar y analizar series temporales. La función de distribución debe ser evaluada

en como una función de n argumentos, por lo que. La representación gráfica es casi imposible

- Las distribuciones marginales y sus funciones de densidad de probabilidad correspondientes, cuando existen, son informativas para examinar el comportamiento marginal de una serie. De este modo, es posible utilizar otras medidas como la media

$$F_t(x) = P(x_t \leq x) \quad f_t(x) = \frac{\partial F_t(x)}{\partial x}$$

- La función de media se define como la esperanza de la variable x_t siempre que esta exista

$$\mu_{x_t} = E(x_t) = \int_{-\infty}^{\infty} x f_t(x) dx$$

- A partir de utilizar el operador de esperanza, es posible obtener la media de varios modelos diferentes

$$v_k = \frac{1}{k} \sum_{j=-k}^k w_{t+j} \Rightarrow E(v_k) = \frac{1}{k} \sum_{j=-k}^k E(w_{t+j}) = 0$$

$$x_t = \phi_0 t + \sum_{j=1}^t w_j \Rightarrow E(x_t) = \phi_0 t + \sum_{j=1}^t E(w_j) = \phi_0 t$$

- La falta de independencia entre dos valores adyacentes x_s y x_t puede ser evaluada numéricamente usando las nociones de covarianza y correlación. Asumiendo que la varianza de x_t es finita, la función de autocovarianza se define como el segundo momento central para toda s y t

$$\gamma_x(s, t) = Cov(x_s, x_t) = E[(x_s - \mu_{x_s})(x_t - \mu_{x_t})]$$

- Debido a la definición, se puede ver como el orden de los momentos no cambia la función de autocovarianza, de modo que $\gamma_x(s, t) = \gamma_x(t, s)$
- La autocovarianza mide la dependencia lineal entre dos puntos en la misma serie observados en momentos diferentes. Series que son muy suaves exhiben funciones de autocovarianzas que se mantienen altas, aunque t y s estén muy lejos en el tiempo, mientras que series menos suaves tienden a exhibir funciones con valores cercanos a cero para grandes separaciones

- Si $\gamma_x(s, t) = 0$, x_s y x_t no están linealmente relacionadas, pero aún puede haber una estructura de dependencia entre ellas. Solo en el caso de normalidad multivariante se tiene que $\gamma_x(s, t) = 0$ implica independencia
- Está claro que, para $s = t$, la función de autocovarianza se convierte a la función de varianza (la cual se asume que es finita)

$$\gamma_x(s, s) = Cov(x_s, x_s) = Var(x_s) = E[(x_s - \mu_{x_s})^2]$$

- Está claro que, para $s = t$, la función de autocovarianza se convierte a la función de varianza (la cual se asume que es finita)
- Una propiedad muy útil de la función de la autocovarianza es la que tiene que ver con variables aleatorias que son combinaciones lineales de otras variables aleatorias
 - Si las variables aleatorias $U = \sum_{j=1}^m a_j X_j$ y $V = \sum_{k=1}^r b_k Y_k$ son combinaciones lineales de las variables aleatorias con varianza finita $\{X_j\}$ y $\{Y_k\}$, respectivamente, entonces se cumple la siguiente identidad:

$$Cov(U, V) = \sum_{j=1}^m \sum_{k=1}^r a_j b_k Cov(X_j, Y_k)$$

- Algunos ejemplos de la función de autocovarianza para algunos modelos anteriormente vistos son los siguientes:
 - La autocovarianza de un ruido blanco se puede expresar de la siguiente manera:

$$Cov(w_t, w_s) = \begin{cases} \sigma_w^2 & \text{if } t = s \\ 0 & \text{if } t \neq s \end{cases}$$

- La autocovarianza de un camino aleatorio se puede expresar de la siguiente manera:

$$Cov(x_t, x_s) = Cov\left(\sum_{j=1}^t w_j, \sum_{k=1}^s w_k\right) = \min(s, t) \sigma_w^2$$

- La autocovarianza de una media móvil se puede expresar de la siguiente manera, en donde se puede ver como la covarianza entre dos puntos va decreciendo a medida que su separación incrementa (solo depende de la separación y no de la

localización de los puntos, lo que se conoce como estacionariedad débil):

$$\gamma_v(s, t) = \text{Cov}(v_s, v_t) = \text{Cov}\left[\frac{1}{k} \sum_{j=-k}^k w_{t+j}, \frac{1}{k} \sum_{j=-k}^k w_{t+j}\right]$$

when $s = t$:

$$\begin{aligned} \gamma_v(t, t) &= \frac{1}{k^2} \text{Cov}\left[\sum_{j=-k}^k w_{t+j}, \sum_{j=-k}^k w_{t+j}\right] = \\ &= \frac{1}{k^2} [\text{Cov}(w_{t-k}, w_{t-k}) + \text{Cov}(w_{t-k+1}, w_{t-k}) + \cdots \\ &\quad + \text{Cov}(w_t, w_t) + \cdots + \text{Cov}(w_{t+k}, w_{t+k})] = \\ &= \frac{1}{k^2} [\sigma_w^2 + 0 + \cdots + \sigma_w^2 + \cdots + 0 + \sigma_w^2] = \frac{2k+1}{k^2} \sigma_w^2 \end{aligned}$$

when $s = t + 1$:

$$\begin{aligned} \gamma_v(t, t+1) &= \frac{1}{k^2} \text{Cov}\left[\sum_{j=-k}^k w_{t+j}, \sum_{j=-k}^k w_{t+1+j}\right] = \\ &= \frac{1}{k^2} [\text{Cov}(w_{t-k}, w_{t-k+1}) + \text{Cov}(w_{t-k+1}, w_{t-k+1}) + \cdots \\ &\quad + \text{Cov}(w_t, w_{t-k}) + \cdots + \text{Cov}(w_{t+1}, w_t)] = \\ &= \frac{1}{k^2} [\sigma_w^2 + 0 + \cdots + \sigma_w^2 + \cdots + 0 + \sigma_w^2] = \frac{2k}{k^2} \sigma_w^2 = \frac{2}{k} \sigma_w^2 \end{aligned}$$

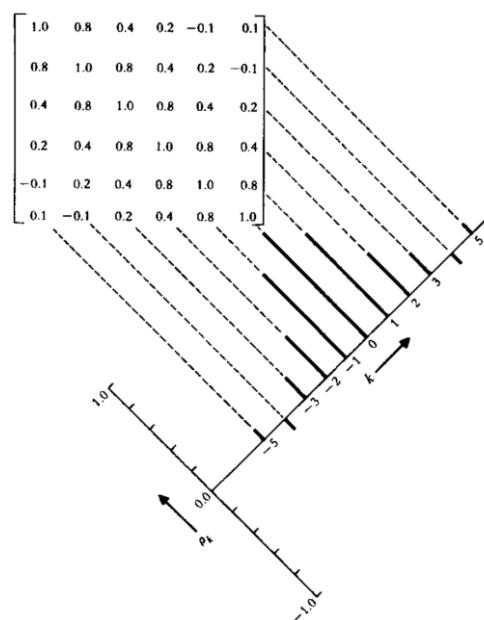
when $|s - t| > k$: $\gamma_v(t, s) = 0$

$$\Rightarrow \gamma_v(t, s) = \begin{cases} \frac{(2k+1)\sigma_w^2}{k^2} & s = t \\ \frac{2\sigma_w^2}{k} & |s-t|=1 \\ \dots & \\ \frac{\sigma_w^2}{k^2} & |s-t|=2k+1 \\ 0 & |s-t| > 2k+1 \end{cases}$$

- La función de autocorrelación (ACF) mide la predictibilidad lineal de una serie en el momento t usando solo el valor x_s , y esta se define de la siguiente manera:

$$\rho(s, t) = \frac{\gamma_x(s, t)}{\sqrt{\gamma_x(s, s)}\sqrt{\gamma_x(t, t)}}$$

- A partir de la desigualdad de Cauchy-Schwarz, es posible demostrar que $\rho(s, t) \in [-1, 1]$
- Si se puede predecir x_t perfectamente a partir de x_s a través de una regresión lineal $x_t = \beta_0 + \beta_1 x_s$, entonces la correlación será 1 si $\beta_1 > 0$ o -1 si $\beta_1 < 0$. Por lo tanto, esta medida se puede interpretar como una medida de la habilidad para predecir la serie en el momento t a través de x_s
- La función de autocorrelación de un proceso generará valores para las diferentes autocorrelaciones que pueda haber (en una matriz simétrica de autocorrelaciones), las cuales se pueden proyectar en un gráfico



- A veces se querrá medir la predictibilidad de otras series y_t a partir de la serie x_s , por lo que, si se asume que ambas series tienen varianzas finitas, se pueden dar las definiciones:

- La función de covarianza cruzada o *cross-covariance* entre dos series y_t y x_s se define de la siguiente manera:

$$\gamma_{xy}(s, t) = Cov(x_s, y_t) = E[(x_s - \mu_{x_s})(y_t - \mu_{y_t})]$$

- La función de correlación cruzada o *cross-correlation* entre dos series y_t y x_s se define de la siguiente manera:

$$\rho_{xy}(s, t) = \frac{\gamma_{xy}(s, t)}{\sqrt{\gamma_x(s, s)}\sqrt{\gamma_y(t, t)}}$$

- Estas ideas se pueden extender al caso de más de una serie, en donde se consideran series temporales multivariante con r componentes $x_{t1}, x_{t2}, \dots, x_{tr}$

- La extensión para la autocovarianza sería la siguiente:

$$\gamma_{jk}(s, t) = E[(x_{sj} - \mu_{sj})(x_{tk} - \mu_{tk})] \quad \text{for } j, k = 1, 2, \dots, r$$

- Las definiciones anteriores de la función de media y autocovarianza son completamente generales, y no se han hecho suposiciones especiales sobre el comportamiento de las series temporales, por lo que ahora se introducirá la noción de estacionariedad

- Una serie temporal estacionaria estrictamente es una en la que su comportamiento probabilístico de cualquier conexión de valores $\{x_{t_1}, x_{t_2}, \dots, x_{t_k}\}$ es idéntico al del conjunto desplazado en el tiempo $\{x_{t_1+h}, x_{t_2+h}, \dots, x_{t_k+h}\}$

$$P(x_{t_1} \leq c_1, x_{t_2} \leq c_2, \dots, x_{t_k} \leq c_k) = P(x_{t_1+h} \leq c_1, x_{t_2+h} \leq c_2, \dots, x_{t_k+h} \leq c_k)$$

- En otras palabras, la distribución de probabilidad es la misma para toda $k = 1, 2, \dots$, todos los puntos en el tiempo $t = t_1, t_2, \dots, t_k$, todos los números c_1, c_2, \dots, c_k y todos los desplazamientos temporales $h = 0, \pm 1, \pm 2, \dots$
- Esta definición de estacionariedad es muy fuerte en la mayoría de aplicaciones. Además, es difícil evaluar la estacionariedad estricta con un solo conjunto de datos
- Si una serie es estrictamente estacionaria, entonces todas las funciones de distribución multivariante para subconjuntos de variables deben de

coincidir con sus contrapartes en el conjunto de datos desplazados para todos los valores del parámetro de desplazamiento h

$$P(x_s \leq c) = P(x_t \leq c)$$

$$P(x_s \leq c_1, x_t \leq c_2) = P(x_{s+h} \leq c_1, x_{t+h} \leq c_2) \quad \text{for } \forall s, t, h$$

- De este modo, la probabilidad de obtener un valor para la serie temporal en un momento t es la misma que la de obtener este mismo valor en un momento s (no depende del tiempo) y la función de media, si esta existe, es la misma para cada momento en el tiempo, por lo que es constante y $\mu_t = \mu_s$
- Además, si la función de autocovarianza existe, entonces la función de la serie es la misma para cualquier s, t y h , por lo que la función de autocovarianza solo depende de la diferencia entre s y t y no en los momentos concretos (depende de la diferencia porque h es la misma para los momentos desplazados en el tiempo)

$$\gamma(s, t) = \gamma(s + h, t + h)$$

- En las definiciones anteriores, las funciones de autocovarianza y covarianza cruzada pueden cambiar cuando uno se mueve a lo largo de la serie porque los valores dependen de s y t , la localización temporal de los puntos
 - En el ejemplo de la autocovarianza para la media móvil, la función de autocovarianza dependía de la separación entre x_s y x_t , $h = |s - t|$, y no en dónde estaban los puntos localizados en el tiempo (la autocovarianza será la misma para una h igual)
 - Esta noción se conoce como estacionariedad débil, la cual es muy importante en el análisis de casos en el que la media es constante. Esta versión no pone condiciones para todas las posibles distribuciones, sino que solo pone condiciones sobre los primeros dos momentos de las series
- Una serie temporal débilmente estacionaria x_t es un proceso de varianza finita tal que la función de media μ_t es constante (no depende del tiempo t) y la función de autocovarianza $\gamma(s, t)$ depende de s y t solo a través de la diferencia $|s - t|$
 - Se usará el término estacionariedad para referirse a la débilmente estacionaria, y si un proceso es estacionario en el sentido estricto, se utiliza el término de estacionariedad estricta

- La estacionariedad requiere regularidad en las funciones de media y autocorrelación, de modo que las cantidades pueden ser estimadas a través de la media
- La estacionariedad estricta también implica que la serie sea estacionaria, pero lo converso no es cierto si no hay condiciones adicionales. No obstante, un caso importante en donde la estacionariedad implica estacionariedad estricta si la serie temporal es gaussiana
- Debido a que la función de media de una serie temporal es independiente del tiempo t se escribe $\mu_t = \mu$. Además, como la función de autocovarianza de una serie estacionaria depende de s y t solo a través de la diferencia $|s - t|$, por lo que si $s = t + h$, entonces se obtiene la siguiente identidad:

$$\gamma(t + h, t) = \text{Cov}(x_{t+h}, x_t) = \text{Cov}(x_h, x_0) = \gamma(h, 0)$$

- Como la diferencia entre los tiempos $t + h$ y t es la misma que la diferencia entre h y 0 . Entonces, la función de autocovarianza de una serie estacionaria no depende del argumento temporal t y se puede sacar el segundo argumento de $\gamma(h, 0)$
- La función autocovarianza de una serie estacionaria se puede escribir de la siguiente manera:

$$\gamma(h) = \text{Cov}(x_{t+h}, x_t) = E[(x_{t+h} - \mu)(x_t - \mu)]$$

- Cuando $h = 0$, entonces se obtiene la varianza de x_t como antes

$$\gamma(0) = \text{Cov}(x_t, x_t) = E[(x_t - \mu)^2]$$

- La función de autocorrelación de una serie estacionaria se puede escribir de la siguiente manera:

$$\rho(h) = \frac{\gamma(t + h, t)}{\sqrt{\gamma(t + h, t + h)}\sqrt{\gamma(t, t)}} = \frac{\gamma(h)}{\gamma(0)}$$

- Algunas propiedades útiles de la función de autocovarianza para series estacionarias son las siguientes:
 - Debido a que $\gamma(h)$ depende solo de h y no de la coordenada temporal t , lo único que importa es la diferencia h y eso hace que $\gamma(h)$ sea la misma mientras que la diferencia entre momentos temporales sea h

$$\text{Cov}(x_{t+h+1}, x_{t+1}) = \text{Cov}(x_{t+h}, x_t) = \text{Cov}(x_{t+h-1}, x_{t-1})$$

$$as \quad (t+h+1) - (t+1) = (t+h-1) - (t-1) = h$$

- La función $\gamma(h)$ es definida no negativa, garantizando que las varianzas de combinaciones lineales de las x_t nunca serán negativas. Por lo tanto, para cualquier $n \geq 1$ y constantes a_1, a_2, \dots, a_n , se cumple la siguiente identidad:

$$0 \leq Var(a_1x_1 + \dots + a_nx_n) = \sum_{j=1}^n \sum_{k=1}^n a_j a_k \gamma(j-k)$$

- Cuando se evalúa la función en $h = 0$, entonces se da que $\gamma(0) = E[(x_t - \mu)^2]$, y la desigualdad de Cauchy-Schwarz implica lo siguiente:

$$|\gamma(h)| \leq \gamma(0)$$

- La función de autocovarianza de una serie estacionaria es simétrica alrededor del origen:

$$\gamma(h) = \gamma(-h) \quad for \quad \forall h$$

$$\begin{aligned} \gamma(h) &= \gamma[(t+h) - t] = Cov(x_{t+h}, x_t) = Cov(x_t, x_{t+h}) = \\ &= \gamma[t - (t+h)] = \gamma(-h) \end{aligned}$$

- Cuando la diferencia entre los momentos temporales no es exactamente h , se puede utilizar el resultado como el nuevo parámetro del cual depende de la función, aunque esta nueva diferencia de un valor negativo (debido a la propiedad anterior)

$$\gamma(h-1) = Cov(x_{t+h+1}, x_t) \neq Cov(x_{t+h}, x_t) = \gamma(h)$$

$$where \quad \gamma(h-1) = \gamma(1-h) \quad \& \quad \gamma(h) = \gamma(-h)$$

- Cuando varias series están disponibles, una noción estacionariedad aún aplica con condiciones adicionales

- Dos series temporales x_t y y_t son conjuntamente estacionarias si cada una de ellas son estacionarias y la función de covarianza cruzada es una función solo del retraso h

$$\gamma_{xy}(h) = Cov(x_{t+h}, y_t) = E[(x_{t+h} - \mu_x)(y_t - \mu_y)]$$

- La función de correlación cruzada de dos series estacionarias conjuntamente x_t y y_t se define de la siguiente manera:

$$\rho_{xy}(h) = \frac{\gamma_{xy}(h)}{\sqrt{\gamma_x(0)\gamma_y(0)}}$$

- La función de correlación cruzada no es simétrica alrededor de cero de manera general, de modo que $\rho_{xy}(h) \neq \rho_{xy}(-h)$. Esto es un concepto importante, por lo que $Cov(x_2, y_1)$ y $Cov(x_1, y_2)$ no necesariamente son lo mismo
- No obstante, se da el caso de que $\rho_{xy}(h) = \rho_{yx}(-h)$ y se puede demostrar utilizando manipulaciones parecidas a las anteriores
- Un ejemplo de estacionariedad conjunta, considerando dos series x_t y y_t , formadas a partir de la suma y la diferencia de dos valores sucesivos de un proceso de ruido blanco en donde w_t son variables aleatorias independientes con media nula y la varianza σ_w^2

$$x_t = w_t + w_{t-1} \quad y_t = w_t - w_{t-1}$$

- A partir de esto se puede demostrar que se dan las siguientes identidades:

$$\begin{aligned}\gamma_x(0) &= \gamma_y(0) = E[(w_t - w_{t-1})^2] = E[w_t^2 - 2w_tw_{t-1} + w_{t-1}^2] \\ &= E(w_t^2) + E(w_{t-1}^2) = 2\sigma_w^2\end{aligned}$$

$$\begin{aligned}\gamma_x(1) &= \gamma_x(-1) = E[(w_{t+1} + w_t)(w_t + w_{t-1})] = \\ &= E(w_{t+1}w_t + w_t^2 + w_tw_{t-1}) = \sigma_w^2\end{aligned}$$

$$\begin{aligned}\gamma_y(1) &= \gamma_y(-1) = E[(w_{t+1} - w_t)(w_t - w_{t-1})] = \\ &= E(w_{t+1}w_t - w_t^2 + w_tw_{t-1}) = \sigma_w^2\end{aligned}$$

- Consecuentemente, se pueden derivar las siguientes propiedades para la función de autocovarianza cruzada:

$$\begin{aligned}\gamma_{xy}(0) &= E[(w_t + w_{t-1})(w_t - w_{t-1})] = \\ &= E(w_tw_t - w_t^2 - w_tw_{t-1}) = 0\end{aligned}$$

$$\begin{aligned}\gamma_{xy}(1) &= E[(w_{t+1} + w_t)(w_t - w_{t-1})] = \\ &= E(w_{t+1}w_t - w_t^2 + w_tw_{t-1}) = \sigma_w^2\end{aligned}$$

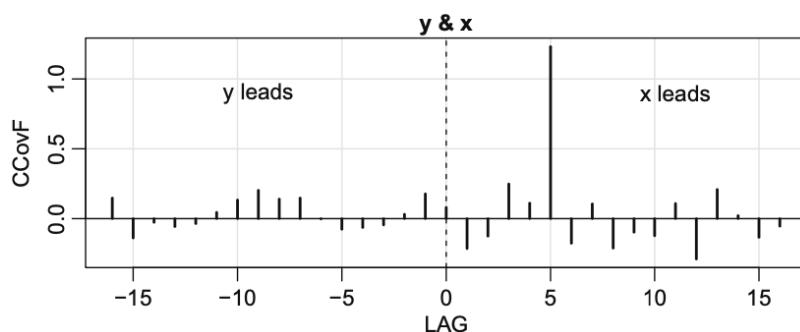
$$\begin{aligned}\gamma_{xy}(-1) &= E[(w_{t-1} + w_{t-2})(w_t - w_{t-1})] = \\ &= E(w_{t-1}w_t - w_{t-1}^2 - w_{t-1}w_{t-2}) = -\sigma_w^2\end{aligned}$$

- Se puede definir la función de autocorrelación en función de los diversos valores de h

$$\rho_{xy}(h) = \begin{cases} 0 & \text{if } h = 0 \\ \frac{1}{2} & \text{if } h = 1 \\ -\frac{1}{2} & \text{if } h = -1 \\ 0 & \text{if } |h| \geq 2 \end{cases}$$

- Como ejemplo de correlación cruzada, se considera el problema de determinar relaciones entre líderes o retrasos entre dos series x_t y y_t
 - Si el modelo $y_t = Ax_{t-l} + w_t$ se mantiene, se dice que la serie x_t lidera y_t para $l > 0$ y se dice que retrasa y_t para $l < 0$. Entonces, los análisis de las relaciones de liderazgo y retraso pueden ser importantes para predecir y_t a partir de x_t
 - Asumiendo que el sonido w_t no está correlacionado con la serie x_t , la covarianza cruzada se puede calcular de la siguiente manera:

$$\begin{aligned}\gamma_{yx}(h) &= Cov(y_{t+h}, x_t) = Cov(Ax_{t+h-l} + w_{t+h}, x_t) = \\ &= Cov(Ax_{t+h-l}, x_t) = ACov(x_{t+h-l}, x_t) = A\gamma_x(h-l)\end{aligned}$$



- El concepto de estacionariedad débil forma la base para muchos análisis de series temporales, las propiedades fundamentales de la media y la autocovarianza se satisfacen para muchos modelos teóricos que generan realizaciones muestrales plausibles

- Un proceso lineal x_t es definido como una combinación lineal de variables w_t que se define de la siguiente manera:

$$x_t = \mu + \sum_{j=-\infty}^{\infty} \psi_j w_{t-j} \quad \sum_{j=-\infty}^{\infty} |\psi_j| < \infty$$

- Para procesos lineales, se puede demostrar que la función de autocovarianza es la siguiente para toda $h \geq 0$:

$$\gamma_x(h) = \sigma_w^2 \sum_{j=-\infty}^{\infty} \psi_{j+h} \psi_j$$

$$\gamma_x(-h) = \sigma_w^2 \sum_{j=-\infty}^{\infty} \psi_{j-h} \psi_j = \sigma_w^2 \sum_{j=-\infty}^{\infty} \psi_{j+h} \psi_j = \gamma_x(h)$$

- Solo es necesario que $\sum_{j=-\infty}^{\infty} \psi_j^2 < \infty$ para que el proceso tenga varianza finita
- El proceso lineal es dependiente del futuro ($j < 0$), del presente ($j = 0$) y del pasado ($j > 0$). Para el propósito de predicción, el modelo dependiente del futuro será inútil. Consecuentemente, uno se enfocará en procesos que no dependan del futuro, llamados modelos causales, en donde $\psi_j = 0$ para $j < 0$
- Aunque la autocorrelación teórica y la función de correlación cruzada son útiles para describir las propiedades de ciertos modelos hipotéticos, la mayoría de análisis se deben realizar con datos muestrales
 - Esta limitación significa que los puntos muestrales x_1, x_2, \dots, x_n solo están disponibles para estimar la función de media, de autocovarianza y de autocorrelación
 - Desde el punto de vista de la estadística clásica, esto crea un problema porque normalmente no se tendrán copias iid de x_t disponibles para estimar las funciones de covarianza y correlación
 - En la situación más común, en donde solo hay una realización, la suposición de estacionariedad se vuelve crítica
 - Se tienen que usar medias para estimar las funciones poblacionales de media y de covarianza. Si una serie temporal es estacionaria, la función de media $\mu_t = \mu$ es constante y se puede estimar la media muestral con la siguiente fórmula:

$$\bar{x} = \frac{1}{n} \sum_{t=1}^n x_t$$

- En este caso, $E(\bar{x}) = \mu$ y el error estándar de este estimador es la raíz cuadrada de $Var(\bar{x})$, que se puede calcular de la siguiente manera:

$$\begin{aligned} Var(\bar{x}) &= Var\left(\frac{1}{n} \sum_{t=1}^n x_t\right) = \frac{1}{n^2} Cov\left(\sum_{t=1}^n x_t, \sum_{s=1}^n x_s\right) = \\ &= \frac{1}{n^2} [n\gamma_x(0) + (n-1)\gamma_x(1) + (n-2)\gamma_x(2) + \dots + \gamma_x(n-1) \\ &\quad + (n-1)\gamma_x(-1) + (n-2)\gamma_x(-2) + \dots + \gamma_x(1-n)] = \\ &= \frac{1}{n} \sum_{h=-n}^n \left(1 - \frac{|h|}{n}\right) \gamma_x(h) \end{aligned}$$

- Como se puede ver, esta fórmula tiene en cuenta que, en caso de dependencia, el error estándar de \bar{x} puede ser mayor o menor que el de un ruido blanco (σ_w^2/n) dependiendo de la naturaleza de la estructura de correlación
- La función de autocovarianza muestral se define de la siguiente manera, de modo que $\hat{\gamma}(h) = \hat{\gamma}(-h)$ para $h = 0, 1, \dots, n-1$:

$$\hat{\gamma}(h) = \frac{1}{n} \sum_{t=1}^{n-h} (x_{t+h} - \bar{x})(x_t - \bar{x})$$

- La suma en esta expresión recorre un rango limitado a $n-h$, ya que x_{t+h} no está disponible para $t+h > n$. Para $h=0$, el estimador se vuelve el estimador de la varianza

$$\hat{\sigma}^2(x_t) = \hat{\gamma}(0) = \frac{1}{n} \sum_{t=1}^n (x_t - \bar{x})^2$$

- Este estimador es preferido al que dividiría por $n-h$ debido a que $\hat{\gamma}(h)$ es una función semidefinida positiva, y concuerda con la función de autocovarianza de un proceso no estacionario (conviene en que es semidefinida positiva), de modo que las varianzas de combinaciones lineales de x_t no serán negativas
- Como la varianza de estas combinaciones lineales no es negativa nunca, entonces el estimador de esa varianza será no negativa y será la siguiente:

$$\hat{\sigma}^2 \left(\sum_{i=1}^n a_i x_i \right) = \sum_{j=1}^n \sum_{k=1}^n a_j a_k \hat{\gamma}(j-k)$$

- La función de autocorrelación muestral se define de la siguiente manera:

$$\hat{\rho}(h) = \frac{\hat{\gamma}(h)}{\hat{\gamma}(0)}$$

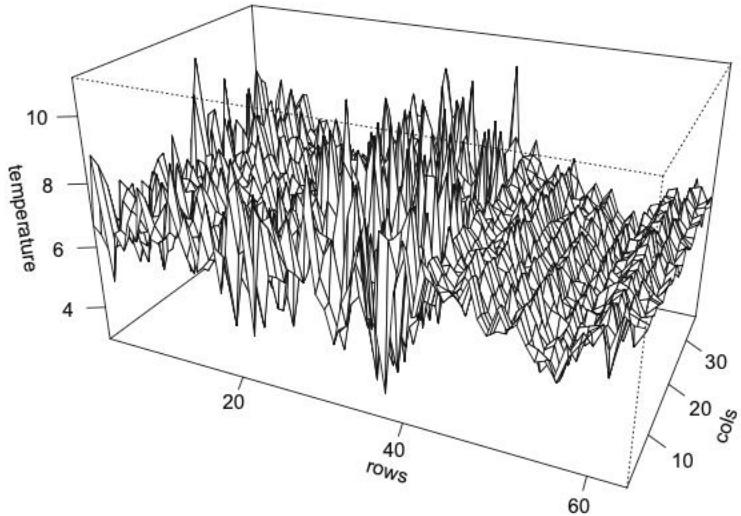
- La función de autocorrelación muestral tiene una distribución muestral que permite evaluar si los datos provienen de una serie completamente aleatoria o de correlaciones que son significativamente estadísticas en algunos retrasos
- Bajo condiciones generales (cuartos momentos finitos), si x_t es ruido blanco, entonces para una n grande, la ACF muestral para $h = 1, 2, \dots, H$ (donde H es fija o arbitraria), se distribuye aproximadamente normal

$$\hat{\rho}_x(h) \sim N\left(0, \frac{1}{n}\right)$$

- Basado en el resultado previo, se pueden construir bandas de confianza para evaluar las autocorrelaciones en sus respectivos retrasos. Para un proceso de ruido blanco, el 95% de las autocorrelaciones muestrales deberían estar dentro de las bandas de confianza

$$IC_{\hat{\rho}_x} = 0 \pm 1.96\sqrt{1/n}$$

- A veces se encuentran situaciones en las que las relaciones entre un número de series temporales medidas conjuntamente son de interés. En esas situaciones, por tanto, es necesario considerar la noción de series temporales de vectores
 - Una serie temporal de vectores $\mathbf{x}_t = (x_{t1}, x_{t2}, \dots, x_{tp})'$ es una serie temporal que contiene p series temporales univariantes como sus componentes



- Se denota el vector columna $p \times 1$ de las series temporales observadas como \mathbf{x}_t . El vector fila \mathbf{x}'_t es la transpuesta de este vector \mathbf{x}_t
- Para el caso estacionario, el vector $p \times 1$ es el vector de medias $\boldsymbol{\mu} = E(\mathbf{x}_t)$ de la forma $\boldsymbol{\mu} = (\mu_{t1}, \mu_{t2}, \dots, \mu_{tp})'$ y la matriz de varianzas y autocovarianzas $p \times p$ puede ser definida, donde los elementos de la matriz $\boldsymbol{\Gamma}(h)$

$$E(\mathbf{x}_t) = \boldsymbol{\mu} = (\mu_{t1}, \mu_{t2}, \dots, \mu_{tp})'$$

$$\boldsymbol{\Gamma}(h) = E[(\mathbf{x}_{t+h} - \boldsymbol{\mu})(\mathbf{x}_t - \boldsymbol{\mu})']$$

$$\gamma_{ij}(h) = E[(\mathbf{x}_{t+h,i} - \mu_i)(\mathbf{x}_{t,j} - \mu_j)] \quad \text{for } i, j = 1, 2, \dots, p$$

- En el caso de no estacionariedad, se obtendrían p valores para la media y la varianza, y $p(p - 1)$ para las covarianzas. Sin embargo, si se asume estacionariedad, el número de parámetros se reduciría ampliamente, dado que se tendría 1 valor para la media y la varianza y $p - 1$ valores para las covarianzas
- Debido $\gamma_{ij}(h) = \gamma_{ij}(-h)$, se puede comprobar que $\boldsymbol{\Gamma}(h) = \boldsymbol{\Gamma}'(h)$
- La matriz de autocovarianza muestral de las series vectoriales \mathbf{x}_t es la matriz de covarianzas cruzadas muestrales, definidas de la siguiente manera:

$$\hat{\boldsymbol{\Gamma}}(h) = n^{-1} \sum_{t=1}^{n-h} (\mathbf{x}_{t+h} - \bar{\mathbf{x}})(\mathbf{x}_t - \bar{\mathbf{x}})'$$

$$\bar{x} = n^{-1} \sum_{t=1}^n x_t$$

- La propiedad simétrica de la autocovarianza teórica se extiende a la autocovarianza muestral, que es definida para los valores negativos tomando la siguiente identidad:

$$\hat{\Gamma}(-h) = \Gamma'(h)$$

- Suponiendo que la posición en el espacio de una unidad experimental puede describirse por dos coordenadas s_1 y s_2 , se puede proceder en estos casos al definir un proceso multidimensional x_s es una función del vector $s = (s_1, s_2, \dots, s_r)'$ de tamaño $r \times 1$ donde s_i denota la coordenada en el índice i
- La función de autocovarianza de un proceso multidimensional estacionario x_s se puede definir como una función del vector de retrasos multidimensional $\mathbf{h} = (h_1, h_2, \dots, h_r)'$ como la siguiente:

$$\gamma(\mathbf{h}) = E[(x_{s+\mathbf{h}} - \mu)(x_s - \mu)] \quad \text{where } \mu = E(x_s)$$

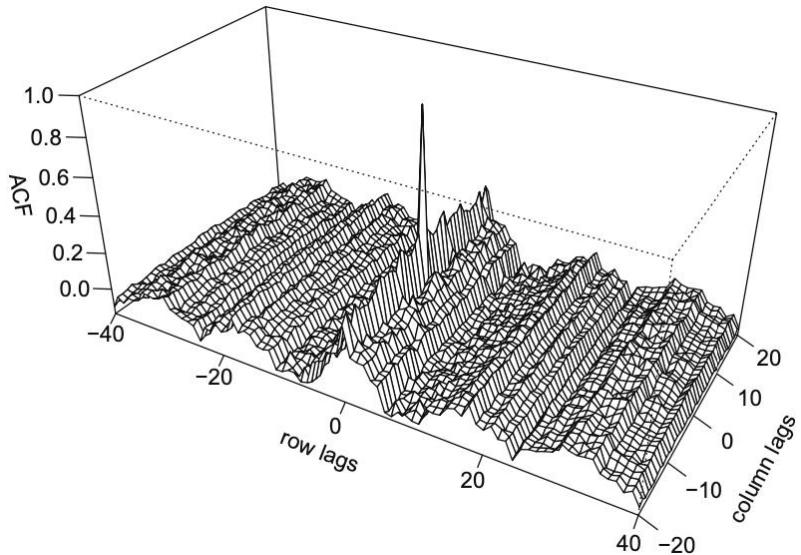
- La función de autocovarianza multidimensional muestral se define de la siguiente manera, donde $s = (s_1, s_2, \dots, s_r)'$:

$$\hat{\gamma}(\mathbf{h}) = (S_1 S_2 \dots S_r)^{-1} \sum_{s_1=1}^{S_1-h} \sum_{s_2=1}^{S_2-h} \dots \sum_{s_r=1}^{S_r-h} (x_{s+h} - \bar{x})(x_s - \bar{x})$$

$$\text{where } \bar{x} = (S_1 S_2 \dots S_r)^{-1} \sum_{s_1=1}^{S_1-h} \sum_{s_2=1}^{S_2-h} \dots \sum_{s_r=1}^{S_r-h} x_{s_1, s_2, \dots, s_r}$$

- La función de autocorrelación muestral se define de manera común:

$$\hat{\rho}(\mathbf{h}) = \frac{\hat{\gamma}(\mathbf{h})}{\hat{\gamma}(0)}$$



Las regresiones con series temporales y análisis exploratorio

- Es posible y útil realizar regresiones lineales múltiples en el contexto de series temporales, de modo que se discuten los modelos y los métodos de selección de modelos e inferencia

- La discusión de la regresión lineal en el contexto de las series temporales se comienza asumiendo que algún resultado o serie dependiente x_t para $t = 1, 2, \dots, n$ se ve influenciada por una colección de posibles insumos o series independientes $z_{t1}, z_{t2}, \dots, z_{tq}$

$$x_t = \beta_0 + \sum_{i=1}^q \beta_i z_{ti} + w_t$$

- Estos insumos se consideran fijos y conocidos para poder aplicar la regresión lineal condicional, aunque se relajará después
- Además, los coeficientes β_i para $i = 0, 1, 2, \dots, q$ se consideran desconocidos y $\{w_t\}$ es un error aleatorio o proceso de ruido consistiendo en variables aleatorias normales $N(0, \sigma_w^2)$ i.i.d
- No obstante, para series temporales no suelen presentar el ruido blanco que se asume aquí, de modo que será necesaria una relajación de esta suposición
- El modelo lineal general presentado puede escribirse de manera más conveniente con una notación más general definiendo los vectores columna $\mathbf{z}_t = (1, z_{t1}, z_{t2}, \dots, z_{tq})'$ y $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \dots, \beta_q)'$

$$x_t = \mathbf{z}_t' \boldsymbol{\beta} + w_t \quad \text{where } w_t \sim \text{iid } N(0, \sigma_w^2)$$

- Los estimadores MCO se pueden estimar a partir del criterio de mínimos cuadrados, de modo que se encuentra un vector $\hat{\beta}$ que minimice la suma cuadrática de errores Q

$$Q = \sum_{t=1}^n w_t^2 = \sum_{t=1}^n (x_t - \mathbf{z}'_t \beta)^2 \Rightarrow \hat{\beta} = \left(\sum_{t=1}^n \mathbf{z}_t \mathbf{z}'_t \right)^{-1} \sum_{t=1}^n \mathbf{z}_t x_t$$

- La suma de errores cuadrados minimizados será la siguiente:

$$SSE = \sum_{t=1}^n (x_t - \mathbf{z}'_t \hat{\beta})^2$$

- Los estimadores de mínimos cuadrados en este contexto cumplen las mismas características que para el caso de datos cruzados, de modo que no tienen sesgo y tienen la mínima varianza dentro de los estimadores lineales
- Si los errores w_t se distribuyen de manera normal, $\hat{\beta}$ también es el estimador máximo verosímil de β y tiene la siguiente matriz de varianzas y covarianzas:

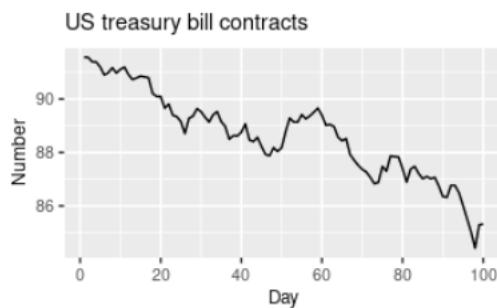
$$Cov(\hat{\beta}) = \sigma_w^2 \left(\sum_{t=1}^n \mathbf{z}_t \mathbf{z}'_t \right)^{-1}$$

- Un estimador no sesgado para la varianza σ_w^2 es el siguiente, en donde el MSE es el error medio cuadrático:

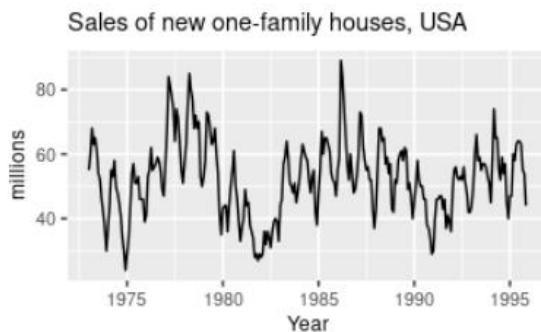
$$s_w^2 = MSE = \frac{SSE}{n - (q + 1)}$$

- En general, es necesario que los datos de una serie temporal sean estacionarios para que hacer la media y análisis derivados de los productos de los valores retrasados en el tiempo sea sensato
 - Con los datos de series temporales, es importante medir la dependencia entre los valores de las series, por lo que se necesita poder estimar las autocorrelaciones con precisión
 - Es difícil medir la dependencia si la estructura de dependencia no es regular o si cambia en cualquier punto en el tiempo. Por lo tanto, para conseguir cualquier análisis estadístico de series de datos temporales, será crucial cumplir con la condición de estacionariedad

- Este no suele ser el caso para muchas series, por lo que se proponen métodos para poder reducir estos efectos de la no estacionariedad para que las propiedades estacionarias de la serie se puedan estudiar
- Al describir una serie temporal, es posible descomponerla en varios elementos que pueden clasificarse en estacionarios y no estacionarios. Los elementos no estacionarios son la tendencia, la estacionalidad y la ciclicidad, mientras que el estacionario se conoce como término irregular
 - Se dice que existe una tendencia cuando hay un crecimiento o decrecimiento a largo plazo en los datos, y esta puede no ser lineal



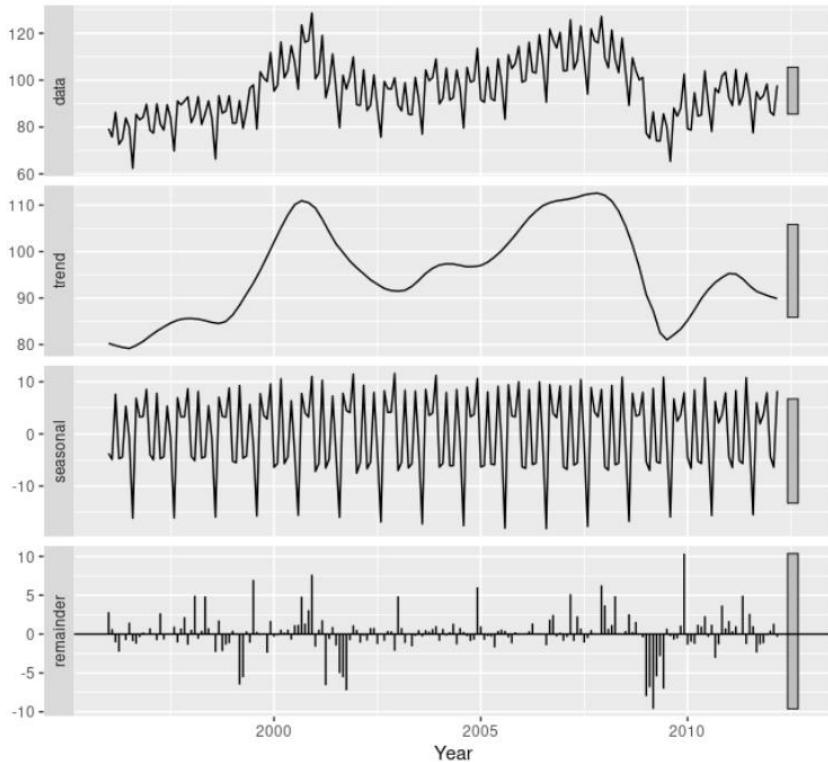
- Un patrón estacional ocurre cuando una serie temporal es afectada por factores estacionales tal como el tiempo del año o el día de la semana. La estacionalidad tiene siempre una frecuencia fija o conocida



- Un ciclo ocurre cuando los datos exhiben subidas y bajadas que no tienen una frecuencia fija. Muchas personas confunden un comportamiento cíclico con un comportamiento estacional, pero si las fluctuaciones no tienen una frecuencia fija, entonces se trata de un ciclo y eso puede afectar a la predicción
- Estos componentes se pueden modelar a través de un modelo que asume una forma funcional de estos factores para x_t . Los dos modelos más utilizados son

$$x_t = \mu_t + c_t + s_t + y_t \quad x_t = \mu_t c_t s_t y_t$$

- En este caso, μ_t es el componente de tendencia, c_t es el componente de ciclos, s_t es el componente de estacionalidad e y_t es el componente estacionario subyacente en la serie temporal (por suposición). El componente c_t es muy difícil de representar debido a que depende de la naturaleza de la serie, por lo que se suele omitir en paquetes de software



- La descomposición aditiva es la más apropiada si la magnitud de las fluctuaciones estacionales o la variación de los datos no varía con el nivel de la serie temporal. Cuando este no es el caso y hay proporcionalidad, entonces una descomposición multiplicativa es mucho más adecuada
- Una alternativa a utilizar la descomposición multiplicativa es transformar la serie hasta que la variación parezca estable en el tiempo. Cuando se usa una transformación logarítmica, es lo mismo usar descomposición aditiva en los datos logarítmicos que usar un modelo multiplicativo en los datos originales

$$x_t = \mu_t s_t y_t$$

$$\Rightarrow \log(x_t) = \log(\mu_t) + \log(s_t) + \log(y_t)$$

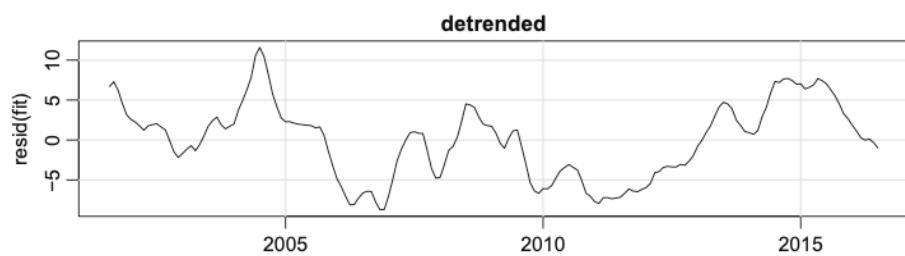
- Antes de poder trabajar con los datos de series temporales y conseguir analizarlos estadísticamente, es necesario hacer un análisis de datos exploratorio y utilizar métodos que permitan trabajar con los datos y la estacionariedad subyacente
 - La forma más fácil de no estacionariedad con la que se puede trabajar es con el modelo de estacionariedad de tendencia en donde el proceso tiene un comportamiento estacionario alrededor de su tendencia

$$x_t = \mu_t + y_t$$

- En este caso, x_t son las observaciones, μ_t denota la tendencia e y_t denota un proceso estacionario
- Normalmente una tendencia fuerte oscurecerá el comportamiento estacionario del proceso y_t . Por lo tanto, hay ventajas al quitar la tendencia como un primer paso en el análisis exploratorio de los datos de series temporales
- Los pasos requeridos son tales que se obtiene un estimador razonable $\hat{\mu}_t$ del componente tendencioso y trabajar con los residuos

$$\hat{y}_t = x_t - \hat{\mu}_t$$

- Suponiendo el modelo anterior, se pueden utilizar varios modelos lineales para poder obtener una estimación $\hat{\mu}_t$



- Para poder hacer una estimación, es posible ajustar un modelo lineal que dependa del tiempo

$$\mu_t = \beta_0 + \beta_1 t$$

- Este modelo puede ser polinómico de un grado $p > 1$, añadiendo complejidad a la tendencia

$$\mu_t = \beta_0 + \sum_{i=1}^p \beta_i t^i$$

- Otro modelo que puede ajustarse a los datos podría ser también un camino aleatorio. De este modo, más que modelar la tendencia como algo dijo, se modela como un componente estocástico con una tasa de deriva

$$\mu_t = \delta + \mu_{t-1} + w_t$$

- En este caso, δ es la tasa de deriva, w_t es ruido blanco y es independiente de y_t
- Si este es el modelo apropiado, entonces diferenciar los datos x_t permite obtener un proceso estacionario

$$\nabla x_t = x_t - x_{t-1} = (\mu_t + y_t) - (\mu_{t-1} + y_{t-1})$$

$$= \delta + w_t + y_t - y_{t-1} = \delta + w_t + z_t$$

- Es posible demostrar que z_t es un proceso estacionario a través de la suposición de estacionariedad de y_t , por lo que $x_t - x_{t-1}$ es estacionario también (al ser combinación lineal de dos procesos estacionarios)

$$E(z_t) = E(y_t) - E(y_{t-1}) = E(y_t) - E(y_t) = 0$$

$$\begin{aligned} \gamma_z(h) &= Cov(z_{t+h}, z_t) = Cov(y_{t+h} - y_{t+h-1}, y_t - y_{t-1}) = \\ &= Cov(y_{t+h}, y_t) - Cov(y_{t+h-1}, y_t) - Cov(y_{t+h}, y_{t-1}) + \\ &\quad + Cov(y_{t+h-1}, y_{t-1}) = 2\gamma_y(h) - \gamma_y(h+1) - \gamma_y(h-1) \end{aligned}$$

$$\text{as } Cov(y_{t+h}, y_t) = Cov(y_{t+h-1}, y_{t-1})$$

- La primera diferencia es un ejemplo de un filtro lineal aplicado a eliminar la tendencia, pero hay otros tipos de filtros que sirven para eliminar otros tipos de fluctuaciones indeseadas
- Una ventaja de la diferenciación sobre la eliminación de la tendencia es que no se necesita estimar parámetros al diferenciar, pero la diferenciación no permite obtener una estimación del proceso estacionario y_t
 - Si una estimación de y_t es esencial, entonces eliminar la tendencia podría ser lo más apropiado, pero si el objetivo es hacer que los datos sean estacionarios, entonces es mejor diferenciar

- La diferenciación también es viable cuando la tendencia es fija por las mismas razones de antes. Para el caso lineal, se obtiene el siguiente resultado:

$$\mu_t = \beta_0 + \beta_1 t$$

$$\nabla x_t = (\mu_t + y_t) - (\mu_{t-1} + y_{t-1}) = \beta_1 + y_t - y_{t-1}$$

- Si la tendencia es de un orden $p > 1$, entonces se tiene que diferenciar p veces para eliminar la tendencia. Para poder definir estas diferencias, es necesario utilizar notación alternativa
 - Es importante no llegar a la sobre-diferenciación, el cual es el caso en el que se diferencia la serie temporal tantas veces que se aumenta la varianza. Si la última diferenciación hecha aumenta la varianza, entonces esta no es necesaria
 - Se define el operador de retroceso o *backshift operator* por $Bx_t = x_{t-1}$ y se extiende a las potencias $B^2x_t = B(Bx_t) = Bx_{t-1} = x_{t-2}$ y así, por lo que $B^kx_t = x_{t-k}$
 - La idea de un operador inverso también se puede dar si se requiere que $B^{-1}B = 1$, de modo que $x_t = B^{-1}Bx_t = B^{-1}x_{t-1}$. Por lo tanto, B^{-1} se conoce como el operador de desplazamiento hacia adelante
 - Además, se puede ver como ∇x_t se puede reescribir como $(1 - B)x_t$ y se puede extender esta noción. Por ejemplo, para la segunda diferencia, se obtiene la siguiente igualdad:
- $$\nabla^2 x_t = (1 - B)^2 x_t = (1 - 2B + B^2)x_t = x_t - 2x_{t-1} + x_{t-2}$$
- $$\nabla^2 x_t = \nabla(\nabla x_t) = \nabla(x_t - x_{t-1}) = (x_t - x_{t-1}) - (x_{t-1} - x_{t-2})$$
- Las diferencias de la orden d se define de $\nabla^d = (1 - B)^d$, donde se puede extender el operador $(1 - B)^d$ algebraicamente para evaluar valores más grandes de d
 - De manera similar, se puede definir la diferenciación estacional con retraso s como $\nabla_s x_t = (1 - B^s)x_t = x_t - x_{t-s}$, el cual es otro filtro lineal
 - Esta diferenciación permitiría eliminar el patrón estacional en los datos y permite definir la media móvil en función de este operador:

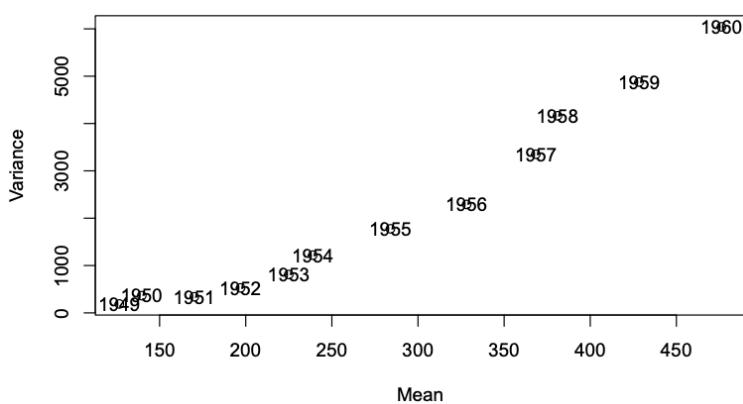
$$\frac{1}{s} \sum_{i=1}^s x_{t-i+1} = \frac{1}{s} (1 + B + \cdots + B^{s-1}) x_t$$

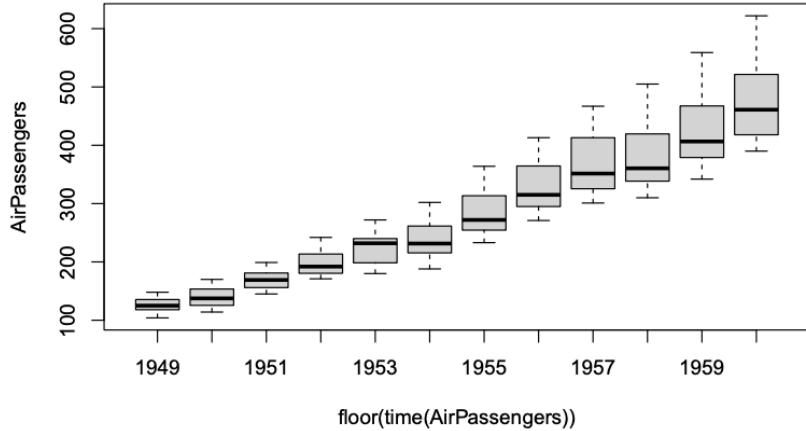
- El diferenciador estacional es equivalente a hacer la primera diferencia de un proceso de media móvil:

$$\nabla_s x_t = (1 - B^s) x_t = (1 - B)(1 + B + \cdots + B^{s-1}) =$$

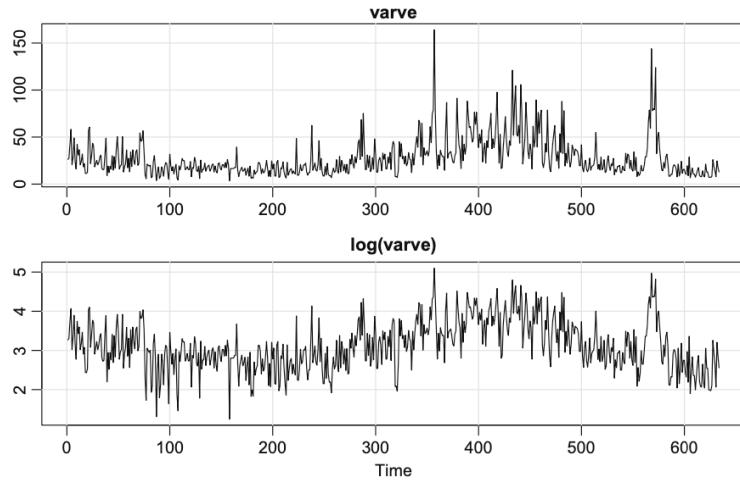
$$= \nabla \left[\sum_{i=1}^s x_{t-i+1} \right]$$

- Una alternativa a diferenciar es utilizar una operación menos severa que aún asume la estacionariedad de la serie temporal subyacente, llamada diferenciación fraccional
 - Esta alternativa extiende la noción del operador a potencias fraccionales $-0.5 < d < 0.5$, que aún define procesos estacionarios
 - Granger, Joyeaux y Hossking introdujeron los procesos de series temporales de memoria larga, que corresponden al caso en donde $0 < d < 0.5$
- Muchas veces hay aberraciones presentes en los datos que contribuyen a la no estacionariedad y a la no linealidad en la serie temporal observada. En tales casos, las transformaciones pueden ser útiles para igualar la variabilidad en la serie temporal
 - Para poder diagnosticar los cambios en la variabilidad se pueden usar métodos gráficos como un gráfico de media contra varianza, en donde se compara la media de diversos grupos de observaciones con su varianza, y el gráfico de cajas, en donde se hace lo mismo, pero a través del rango intercuartílico





- Una transformación particularmente útil es $y_t = \log(x_t)$, que tiende a suprimir fluctuaciones grandes que ocurren sobre porciones de la serie en la que los valores subyacentes son grandes

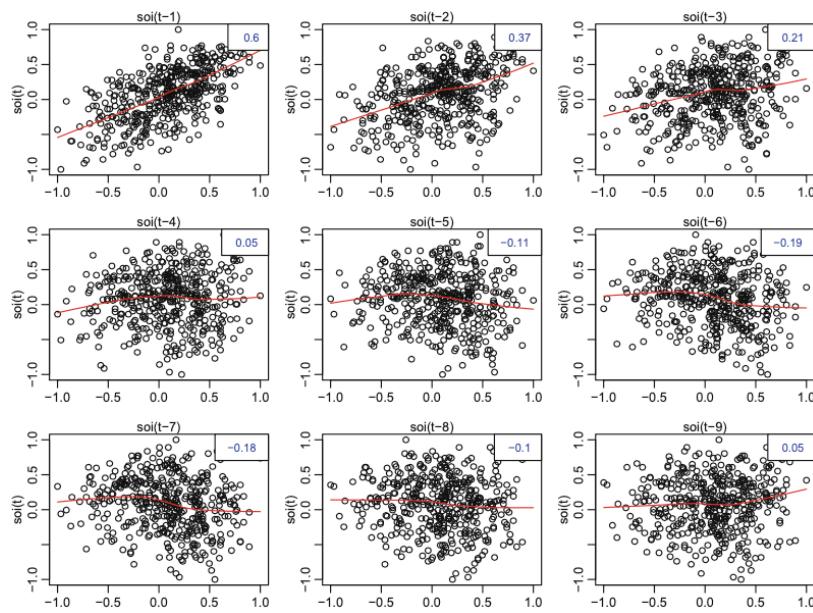


- Otras posibilidades son las transformaciones de potencias en la familia de Box-Cox, las cuales tienen la siguiente forma:

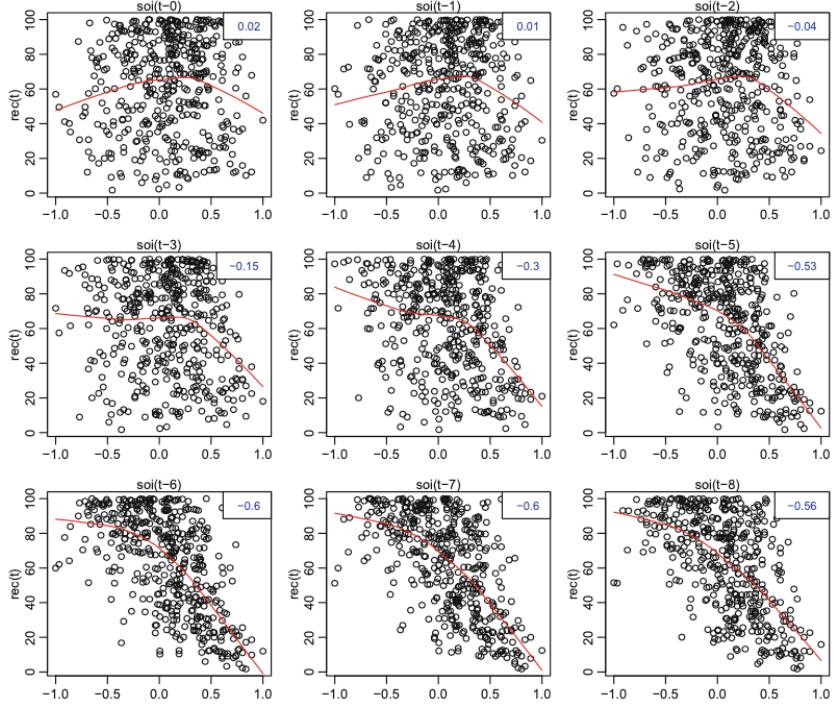
$$y_t = \begin{cases} (x_t^\lambda - 1)/\lambda & \text{if } \lambda \neq 0 \\ \log(x_t) & \text{if } \lambda = 0 \end{cases} \quad \text{where } \lambda \in [-1, 2]$$

- Existen métodos para escoger la potencia λ , tales como $\lambda = 0$ cuando la variabilidad es mayor para valores más elevados de x_t o $\lambda = -1$ cuando ocurre el caso contrario (respecto al eje vertical x_t , no al tiempo t , por lo que la orientación de la serie no importa, solo el nivel)
- Cuando una serie tiene valores negativos, y se quiere aplicar una transformación de Box-Cox, a veces no se puede calcular la transformación para valores negativos. Por lo tanto, se puede sumar 1 más el valor mínimo (en valor absoluto) de la serie con tal de que no hayan valores negativos y aplicar la transformación

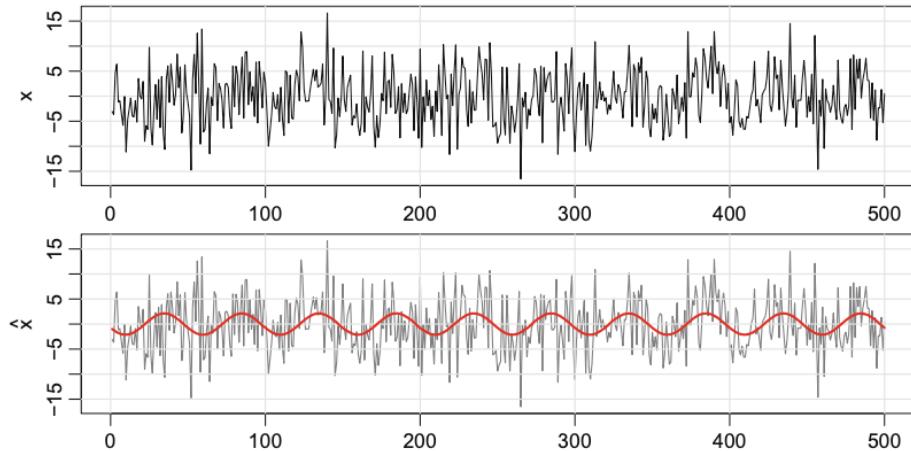
- También se pueden considerar unas técnicas de procesamiento de datos que se usa para poder visualizar las relaciones entre las series en sus diferentes retrasos, tales como las matrices de gráficos de dispersión
 - En la definición de la ACF, uno se interesa esencialmente en las relaciones entre x_t y x_{t-h} , y esta función dice si existe una relación lineal sustancial entre las series y sus valores retrasados
 - La ACF da un perfil de la correlación lineal en todos los posibles retrasos y muestra que valores de h derivan a una mejor predicción



- La restricción de esta idea a la predicción lineal, no obstante, puede enmascarar posibles correlaciones no lineales entre los valores actuales x_t y sus retrasos x_{t-h}



- Como herramienta exploratoria final, se puede discutir la evaluación del comportamiento periódico de una serie temporal utilizando el análisis de regresiones



- Hay muchas series (sobre todo en series temporales físicas) que tienen una periodicidad clara, por lo que para ello se pueden utilizar regresiones lineales con tal de estimar parámetros que tengan contenidas información sobre la periodicidad
- Una vez vistas las técnicas de análisis exploratorio de los datos, es posible hacer un cuadro resumen para poder analizar la estacionariedad subyacente en los datos

Non stationarity cause	Transformation
Linear deterministic trend	Differencing $(1 - B)$. The mean of the series obtained is the (constant) slope of the trend
Deterministic d order polynomial trend	Differencing $(1 - B)^d$
Stochastic trend	Differencing $(1 - B)^d$ until the series becomes stationary (make sure there is no overdifferentiating)
Non constant mean	Differencing $(1 - B)$
S order seasonality	Differencing $(1 - B^s)$. It also removes a linear deterministic trend.
Seasonality, constant variance and linear trend with variable slope (Stochastic trend)	Differencing $(1 - B^s) \cdot (1 - B)$
Non constant variance	Box-Cox Transformation (A particular case is the logarithm transformation, used when the relation between the variance and the mean is linear in consecutive years of the series).

- Primero se intenta eliminar el componente tendencioso de las series, y una vez eliminado, se pasa a transformar la varianza en una varianza constante. Cuando ambas cosas se han realizado, se intenta eliminar el componente estacional, de modo que, si se han solucionado estas tres cuestiones, se llega a una serie estacionaria
- Anteriormente se ha introducido el concepto de filtro o suavizado de una serie temporal, sobre todo en el contexto de la media móvil para suavizar el ruido blanco. Este método es útil para poder descubrir ciertos aspectos de las series temporales como tendencias a largo plazo o componentes estacionales
 - En particular, si x_t representa las observaciones, entonces m_t , definido de la siguiente manera, es una media móvil simétrica de los datos:

$$m_t = \sum_{j=-k}^k a_j x_{t-j}$$

- En este caso, se tienen que cumplir las siguientes propiedades para los coeficientes:

$$a_j = a_{-j} \geq 0 \quad \& \quad \sum_{j=-k}^k a_j = 1$$

- Es posible obtener un ajuste mucho más suave si se utilizan *kernels* en las ponderaciones. El suavizado de *kernel* es un suavizante de media móvil que utiliza una función de ponderación $K(\cdot)$ para promediar las observaciones

$$m_t = \sum_{i=1}^n w_i(t) x_i \quad \text{where} \quad w_i(t) = \frac{K\left(\frac{t-i}{b}\right)}{\sum_{j=1}^n K\left(\frac{t-j}{b}\right)}$$

- Este estimador normalmente se llama estimador de Nadaraya-Watson, y normalmente se utiliza el *kernel* normal

$$K(z) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right)$$

Los modelos ARIMA: los modelos estacionarios

- La regresión clásica normalmente es insuficiente para explicar todas las dinámicas interesantes de las series temporales
 - La introducción de la correlación generada por el uso de valores retrasados y el uso de series temporales no estacionarias justifican el uso de modelos más flexibles como los modelos ARIMA
 - Dentro de este tipo de modelos se pueden encontrar modelos tan importantes como los modelos autorregresivos, los modelos de media móvil y los modelos ARMA
 - Estos modelos permiten capturar estructuras adicionales que no han sido previamente capturadas por los modelos de regresión
 - A partir de estos modelos se han podido desarrollar metodologías estadísticas para la identificación y el análisis
 - La metodología de Box-Jenkins permite identificar modelos ARIMA
 - Existen técnicas para la estimación paramétrica y la predicción especializadas para series temporales
- Los modelos autorregresivos se basan en la idea de que el valor actual de la serie temporal x_t se puede expresar como función de p valores pasados $x_{t-1}, x_{t-2}, \dots, x_{t-p}$ donde p determina el número de pasos en el pasado necesarios para predecir el valor actual
 - Un modelo autorregresivo de orden p , abreviado como $AR(p)$ es un modelo de la siguiente forma:

$$x_t = \phi_1 x_{t-1} + \phi_2 x_{t-2} + \dots + \phi_p x_{t-p} + w_t$$

$$\text{where } w_t \sim \text{wn}(0, \sigma_w^2)$$

- En este modelo, se asume que x_t es una serie estacionaria y que ϕ_i para $i = 1, 2, \dots, p$ son constantes y $\phi_p \neq 0$

- Se asume que la media de x_t es nula, aunque si la media es $\mu \neq 0$, se tiene que reemplazar x_t por $x_t - \mu$ o añadir un término $\alpha = (1 - \phi_1 - \phi_2 - \dots - \phi_p)$

$$x_t - \mu = \phi_1(x_{t-1} - \mu) + \dots + \phi_p(x_{t-p} - \mu) + w_t$$

$$x_t = \alpha + \phi_1 x_{t-1} + \phi_2 x_{t-2} + \dots + \phi_p x_{t-p} + w_t$$

- Existen problemas técnicos a la hora de utilizar el modelo de autorregresión anterior porque los regresores $x_{t-1}, x_{t-2}, \dots, x_{t-p}$ son componentes aleatorios y z_t es un componente fijo por suposición
- Este modelo se puede expresar de manera equivalente a través del operador de retraso

$$(1 - \phi_1 B - \phi_2 B^2 - \dots - \phi_p B^p)x_t = w_t$$

- Es posible expresar el modelo de manera mucho más concisa definiendo el operador autorregresivo. Este operador se define de la siguiente manera:

$$\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \dots - \phi_p B^p$$

$$\Rightarrow \phi(B)x_t = w_t$$

- Para poder iniciarse en la investigación de los modelos AR se considera el modelo $AR(1)$, dado por $x_t = \phi_1 x_{t-1} + w_t$
- Iterando el modelo hacia atrás k veces, se puede ver como, si se iterara hacia atrás infinitamente y se suponiera que $|\phi_1| < 1$ y $\sup Var(x_t) < \infty$, se puede expresar x_t para como un proceso lineal del siguiente tipo:

$$x_t = \phi_1(\phi_1 x_{t-2} + w_t) + w_t = \phi_1(\phi_1(\dots) + w_t) + w_t =$$

$$= \phi_1^k x_{t-k} + \sum_{j=0}^{k-1} \phi_1^j w_{t-j}$$

$$\Rightarrow x_t = \sum_{j=0}^{\infty} \phi_1^j w_{t-j} \quad \text{when } |\phi_1| < 1 \quad \& \quad \sup Var(x_t) < \infty,$$

- Esta última representación se llama solución estacionaria del modelo

$$x_t = \sum_{j=0}^{\infty} \phi_1^j w_{t-j} \Rightarrow \sum_{j=0}^{\infty} \phi_1^j w_{t-j} = \phi_1 \left(\sum_{k=0}^{\infty} \phi_1^k w_{t-1-k} \right) + w_t$$

- El proceso $AR(1)$ definido por $x_t = \sum_{j=0}^{\infty} \phi_1^j w_{t-j}$ es un proceso estacionario con la siguiente media y autocovarianza

$$E(x_t) = \sum_{j=0}^{\infty} \phi_1^j E(w_{t-j}) = 0$$

$$\begin{aligned} \gamma(h) &= Cov(x_{t+h}, x_t) = E \left[\left(\sum_{j=0}^{\infty} \phi_1^j w_{t+h-j} \right) \left(\sum_{k=0}^{\infty} \phi_1^k w_{t-k} \right) \right] = \\ &= E[(w_{t+h} + \cdots + \phi_1^h w_t + \phi_1^{h+1} w_{t-1} + \cdots)(w_t + \phi_1 w_{t-1} + \cdots)] = \\ &= E(w_{t+h} w_t) + \cdots + \phi_1^h E(w_t^2) + \phi_1^{h+1} E(w_{t-1} w_t) + \cdots = \\ &= \sigma^2 \sum_{j=1}^{\infty} \phi_1^{h+j} \phi_1^j = \sigma^2 \phi_1^h \sum_{j=1}^{\infty} (\phi_1^2)^j = \frac{\sigma^2 \phi_1^h}{1 - \phi_1^2} \quad \text{for } h \geq 0 \end{aligned}$$

- Otra manera sencilla de poder encontrar la función de autocovarianza es a través de cambiar $Cov(x_{t+h}, x_t)$ por $Cov(x_t, x_{t-h})$ y desarrollar h veces el modelo completo para las x_{t-h}

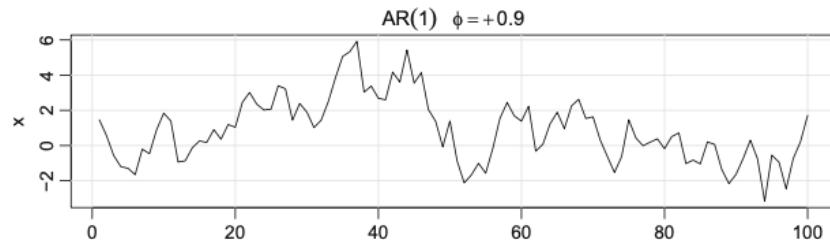
$$\begin{aligned} \gamma(h) &= Cov(x_{t+h}, x_t) = Cov(x_t, x_{t-h}) = \\ &= Cov(\phi_1 x_{t-1} + w_t, x_{t-h}) = \\ &= Cov(\phi_1(\phi_1(\phi_1(\dots) + w_{t-2}) + w_{t-1}) + w_t, x_{t-h}) \end{aligned}$$

- Debido a que $\gamma(h) = \gamma(-h)$, solo es necesario mostrar la función de autocovarianza para $h \geq 0$. A partir de esta, también se puede obtener la función de autocorrelación y una forma recursiva de esta

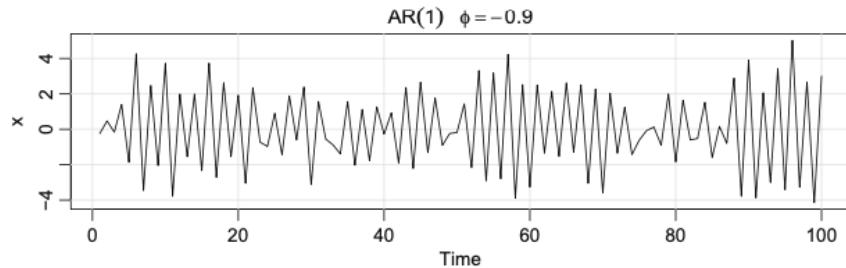
$$\begin{aligned} \gamma(0) &= \frac{\sigma^2}{1 - \phi_1^2} \Rightarrow \rho(h) = \frac{\gamma(h)}{\gamma(0)} = \phi_1^h \\ &\Rightarrow \rho(h) = \phi_1 \rho(h-1) \end{aligned}$$

- Otra de las cosas más importantes al estudiar los procesos autorregresivos es el estudio del camino que sigue una muestra. En este caso, también es útil introducirse a través del $AR(1)$ con $|\phi_1| < 1$

- La autocorrelación será positiva para toda $h > 0$ si $\phi_1 > 0$, mientras que si $\phi_1 < 0$, entonces la autocorrelación será negativa para h impares pero positiva para h pares
- En el caso en que ϕ_1 esté muy cerca de 1, la autocorrelación entre observaciones será positiva, por lo que observaciones contiguas en el tiempo tenderán a estar cerca, haciendo que el camino muestral sea muy suave



- En el caso en que ϕ_1 esté muy cerca de -1, la autocorrelación entre observaciones será negativa para retrasos impares pero positiva para retrasos pares, por lo que los signos de las observaciones irán cambiando y habrá más distancia entre observaciones contiguas (aunque no más variabilidad), haciendo que el camino muestral sea áspero o duro



- Un proceso $AR(1)$ con $|\phi_1| > 1$ se denomina proceso explosivo, dado que sus valores se vuelven rápidamente grandes (en magnitud)
 - Debido a que $|\phi_1|^j$ incrementa sin una cota superior cuando $j \rightarrow \infty$, la serie $\sum_{j=0}^{k-1} \phi_1^j w_{t-j}$ no convergerá en media cuadrática cuando $k \rightarrow \infty$, de modo que la intuición anterior para encontrar una expresión de x_t en el límite no sirve
 - No obstante, es posible modificar este argumento anterior para poder obtener un modelo estacionario. Escribiendo $x_{t+1} = \phi_1 x_t + w_{t+1}$, por lo que se pueden derivar las siguientes igualdades:

$$x_{t+1} = \phi_1 x_t + w_{t+1} \Rightarrow x_t = \phi_1^{-1} x_{t+1} - \phi_1^{-1} w_{t+1}$$

$$\Rightarrow x_t = \phi_1^{-1}(\phi_1^{-1}(\dots) - \phi_1^{-1}w_{t+2}) - \phi_1^{-1}w_{t+1}$$

$$\begin{aligned}\Rightarrow x_t &= \phi_1^{-k}x_{t+k} + \sum_{j=0}^{k-1} \phi_1^{-j}w_{t+j} \\ \Rightarrow x_t &= -\sum_{j=0}^{\infty} \phi_1^{-j}w_{t+j}\end{aligned}$$

- En este caso, el modelo es estacionario y tiene la forma común de un modelo $AR(1)$, pero es inútil porque es necesario saber el futuro para poder predecir el futuro. Cuando un proceso no depende del futuro, tal como el modelo $AR(1)$ cuando $|\phi| < 1$, este se denomina proceso causal
- Excluir los modelos explosivos de la consideración no es un problema debido a que todos estos modelos tienen una contraparte causal, tal y como se ha visto en el caso del $AR(1)$ explosivo
 - Si $x_t = \phi_1 x_{t-1} + w_t$, donde $|\phi_1| > 1$ y $w_t \sim iid N(0, \sigma^2)$, entonces este proceso no causal estacionario gaussiano con las siguientes medias y varianzas:

$$E(x_t) = 0$$

$$\begin{aligned}\gamma_x(h) &= Cov(x_{t+h}, x_t) = \\ &= E \left[\left(-\sum_{j=0}^{\infty} \phi_1^{-j}w_{t+h+j} \right) \left(-\sum_{k=0}^{\infty} \phi_1^{-k}w_{t+k} \right) \right] = \sigma_w^2 \frac{\phi^{-2} \phi^{-h}}{1 - \phi^{-2}}\end{aligned}$$

- Por lo tanto, el proceso causal de este proceso explosivo se define con la siguiente fórmula, la cual es estocásticamente equivalente al proceso de x_t (todas las distribuciones finitas del proceso son las mismas):

$$y_t = \phi_1^{-1}y_t + v_t \quad \text{where } v_t \sim iid N(0, \sigma_w^2 \phi_1^{-2})$$

- Esto se extiende a procesos autorregresivos de mayor orden, cuyas demostraciones se enseñarán más adelante
- La técnica de iterar hacia atrás funciona bien para tener una idea de como funcionan los modelos autorregresivos a partir de $p = 1$, pero no para $p > 1$, por lo que se considera la técnica general de coeficientes coincidentes

- Considerando un modelo $AR(1)$ con $\phi(B) = 1 - \phi_1 B$ y $|\phi_1| < 1$, y escribiendo $\psi(B) = \sum_{j=0}^{\infty} \psi_j B^j$, es posible obtener la siguiente igualdad:

$$\phi(B)x_t = w_t \quad \& \quad \psi(B)w_t = x_t \text{ iff } \psi_j = \phi_1^j$$

$$\Rightarrow \phi(B)\psi(B)w_t = w_t$$

- Los coeficientes de B en la parte izquierda de la igualdad deben ser iguales a los de la parte derecha, por lo que se establece la siguiente identidad:

$$(1 - \phi_1 B)(1 + \psi_1 B + \psi_2 B^2 + \dots + \psi_j B^j + \dots) = 1$$

$$\Rightarrow 1 + (\psi_1 - \phi_1)B + \dots + (\psi_j - \psi_{j-1}\phi_1)B^j + \dots = 1$$

$$\text{where } \psi_0 = 1$$

- En este caso, para $j = 1, 2, \dots$ el coeficiente de B^j debería ser nulo porque $(\psi_1 - \phi_1)B + \dots + (\psi_j - \psi_{j-1}\phi_1)B^j + \dots$ tiene que ser un cero. A partir de esto, es posible obtener las siguientes igualdades para cada coeficiente:

$$\begin{cases} \psi_1 - \phi_1 = 0 \\ \dots \\ \psi_j - \psi_{j-1}\phi_1 = 0 \\ \dots \end{cases} \Rightarrow \psi_j = \phi_1^j \text{ for } j = 1, 2, \dots$$

- Otra manera de poder pensar en estos coeficientes es a través de considerar un operador inverso para el operador $\phi(B)$ y utilizar polinomios

- Multiplicando $\phi^{-1}(B)$ en ambos lados de $\phi(B)x_t = w_t$ se puede obtener la siguiente igualdad, obteniendo una forma conocida como la forma $MA(\infty)$:

$$\phi^{-1}(B)\phi(B)x_t = \phi^{-1}(B)w_t$$

$$x_t = \phi^{-1}(B)w_t$$

- Debido a que $\phi^{-1}(B) = 1 + \phi_1 B + \phi_1^2 B^2 + \dots + \phi_1^j B^j + \dots$, entonces $\phi^{-1}(B) = \psi(B)$. Esto hace que trabajar con este operador sea igual a trabajar con polinomios

$$\phi^{-1}(B) = \frac{1}{1 - \phi_1 B} = 1 + \phi_1 B + \phi_1^2 B^2 + \dots + \phi_1^j B^j + \dots$$

- Si se considera que $\phi(B) = 1 - \phi_1 z$, en donde z es un número complejo y $|\phi_1| < 1$, entonces los coeficientes de B^j en $\phi^{-1}(B)$ son los mismos que los de z^j en $\phi^{-1}(z)$ (se puede tratar al operador $\phi^{-1}(z)$ como un número complejo)

$$\phi^{-1}(z) = \frac{1}{1 - \phi_1 z} = 1 + \phi_1 z + \phi_1^2 z^2 + \cdots + \phi_1^j z^j + \cdots$$

- Como alternativa a la representación autorregresiva en la que x_t se combina linealmente por suposición, el modelo de media móvil de orden q , abreviado como $MA(q)$, asume que el ruido blanco w_t es una combinación lineal para formar los datos observados
 - El modelo de media móvil de orden q , denotado como $MA(q)$, se define de la siguiente manera, en donde $w_t \sim wn(0, \sigma_w^2)$ y $\theta_1, \theta_2, \dots, \theta_q$ (con $\theta_q \neq 0$) son constantes:

$$x_t = w_t + \theta_1 w_{t-1} + \theta_2 w_{t-2} + \cdots + \theta_q w_{t-q}$$

- El sistema es el mismo que la media móvil infinita, definido como el proceso lineal anteriormente visto, en donde $\psi_0 = 1$, $\psi_j = \theta_j$ para $j = 1, 2, \dots, q$ y $\psi_j = 0$ para $j > q$
- A diferencia del modelo autorregresivo, este modelo es estacionario para cualquier valor de $\theta_1, \theta_2, \dots, \theta_q$
- Este modelo se puede expresar de manera equivalente a través del operador de media móvil

$$x_t = (1 + \theta_1 B + \theta_2 B^2 + \cdots + \theta_q B^q)w_t$$

- Es posible expresar el modelo de manera mucho más concisa definiendo el operador de media móvil. Este operador se define de la siguiente manera:

$$\theta(B) = 1 + \theta_1 B + \theta_2 B^2 + \cdots + \theta_q B^q$$

$$\Rightarrow x_t = \theta(B)w_t$$

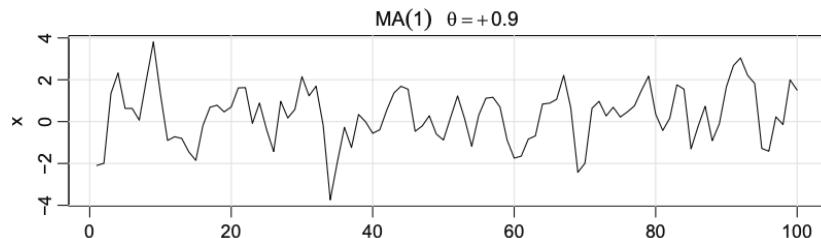
- Para poder iniciarse en la investigación de los modelos AR se considera el modelo $MA(1)$, dado por $x_t = w_t + \theta_1 w_{t-1}$
 - La esperanza, la función de autocovarianza y la de autocorrelación vienen dadas por las siguientes identidades:

$$E(x_t) = E(w_t + \theta_1 w_{t-1}) = E(w_t) + \theta_1 E(w_{t-1}) = 0$$

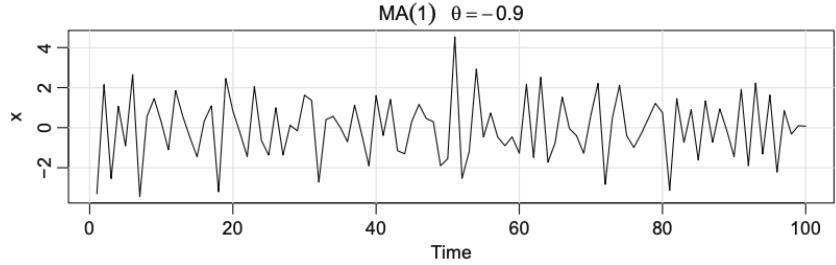
$$\begin{aligned}
\gamma_x(h) &= E(x_{t+h}x_t) = E[(w_{t+h} + \theta_1 w_{t+h-1})(w_t + \theta_1 w_{t-1})] = \\
&= E(w_{t+h}w_t) + \theta_1 E(w_{t+h-1}w_t) + \theta_1 E(w_{t+h}w_{t-1}) + \theta_1^2 E(w_{t-1}w_{t+h-1}) = \\
&= \begin{cases} \sigma_w^2(1 + \theta_1^2) & \text{if } h = 0 \\ \theta_1 \sigma_w^2 & \text{if } h = 1 \\ 0 & \text{if } h > 1 \end{cases}
\end{aligned}$$

$$\rho(h) = \frac{\gamma_x(h)}{\gamma_x(0)} = \begin{cases} 1 & \text{if } h = 0 \\ \theta_1/(1 + \theta_1^2) & \text{if } h = 1 \\ 0 & \text{if } h > 1 \end{cases}$$

- En este caso, $|\rho(1)| \leq 1/2$ para cualquier valor de θ_1 y x_t está correlacionado con x_{t-1} pero no con cualquier otro retraso del proceso, lo cual se diferencia del proceso autorregresivo, en donde ninguna correlación es nula entre x_t y x_{t-k}
- Debido a la no correlación entre los ruidos blancos y la forma funcional no recursiva (a diferencia de con el modelo $AR(1)$), la manera mostrada es la más sencilla para encontrar la función de autocovarianza
- El comportamiento del camino muestral de los MA es prácticamente el mismo que para los procesos autorregresivos
 - La única diferencia es que la autocorrelación para $\rho(1)$ depende completamente del signo de θ_1 y no de la observación, de modo que si $\theta_1 > 0$, entonces la autocorrelación en $h = 1$ será positiva, mientras que si $\theta_1 < 0$ la autocorrelación en $h = 1$ será negativa
 - Cuando θ_1 está cerca de 1, se puede ver como el camino muestral será suave



- Cuando θ_1 está cerca de -1, se puede ver como el camino muestral será áspero o duro



- Además, es posible comprobar como un proceso MA no es único cuando w_t sigue una distribución normal, dado que para cualquier $\theta \neq 0$ existe un coeficiente $1/\theta$ que permite obtener un proceso MA estocásticamente equivalente
 - En este caso, si se utiliza $1/\theta$, es posible comprobar que los siguientes procesos son equivalentes:

$$x_t = w_t + \frac{1}{\theta}w_{t-1} \quad \text{where} \quad w_t \sim \text{iid } N(0, \theta^2 \sigma_w^2)$$

$$y_t = v_t + \theta v_{t-1} \quad \text{where} \quad v_t \sim \text{iid } N(0, \sigma_v^2)$$

- Uno solo puede observar las series temporales, x_t o y_t , y no el ruido w_t o v_t , por lo que no se puede distinguir cuál es el proceso a partir de los datos. Esto hace que se tenga que escoger solo uno de ellos
- Para poder descubrir qué modelo de estos tiene una representación infinita (un proceso invertible), se escribe el modelo $MA(1)$ como $w_t = -\theta w_{t-1} + x_t$ y se obtiene una serie de potencias para w_t que es equivalente a la representación infinita de AR. En este caso, por tanto, se escoge aquel modelo para el cuál exista esta representación (que el coeficiente tenga valor absoluto menor a uno)

$$w_t = -\theta(-\theta(\dots) + x_{t-1}) + x_t = \sum_{j=0}^{\infty} (-\theta)^j x_{t-j} = \sum_{j=0}^{\infty} (-1)^j \theta^j x_{t-j}$$

$$\text{if } |\theta| < 1$$

- En el caso del AR, el polinomio $\theta(z)$, correspondiente al operador de la media móvil es útil para explicar algunas propiedades
 - Considerando un modelo $MA(1)$, es posible expresar este a través de considerar un operador inverso $\pi(B) = \theta^{-1}(B)$, para el operador de la media móvil $\theta(B) = 1 + \theta B$

$$x_t = \theta(B)w_t \Rightarrow \theta^{-1}(B)x_t = \theta^{-1}(B)\theta(B)w_t$$

$$\Rightarrow \theta^{-1}(B)x_t = w_t$$

- Si $|\theta| < 1$, entonces se puede escribir el modelo como $\pi(B)x_t = w_t$, en donde $\pi(B) = \theta^{-1}(B) = 1/(1 + \theta B)$, cuya forma se conoce como $AR(\infty)$. Siendo $\theta(z) = 1 - \theta z$, entonces se pueden obtener las siguientes identidades:

$$\pi(z) = \theta^{-1}(z) = \frac{1}{1 + \theta z} = 1 - \theta z - \theta^2 z^2 - \cdots - \theta^q z^q - \cdots$$

$$\Rightarrow \pi(B) = \sum_{j=0}^{\infty} (-\theta)^j B^j$$

- A partir del desarrollo teórico anterior es posible desarrollar una teoría más general de los modelos AR, MA y de los modelos autorregresivos con media móvil o ARMA, todos para series estacionarias
 - Una serie temporal $\{x_t; t = 0, \pm 1, \pm 2, \dots\}$ es $ARMA(p, q)$ si esta es estacionaria y x_t se puede expresar de la siguiente manera, con $\phi_p \neq 0$, $\theta_q \neq 0$ y $\sigma_w^2 > 0$:

$$x_t = \phi_1 x_{t-1} + \cdots + \phi_p x_{t-p} + w_t + \theta_1 w_{t-1} + \cdots + \theta_q w_{t-q}$$

- Los parámetros p y q se denominan órdenes autorregresivos y de media móvil, respectivamente
- Si x_t tiene una media $\mu \neq 0$, entonces se fija $\alpha = (1 - \phi_1 - \cdots - \phi_p)$ y se escribe el modelo de la siguiente manera:

$$x_t = \alpha + \phi_1 x_{t-1} + \cdots + \phi_p x_{t-p} + w_t + \theta_1 w_{t-1} + \cdots + \theta_q w_{t-q}$$

- También es posible eliminar la media de la serie temporal con tal de aplicar el modelo original:

$$(x_t - \mu) = \phi_1(x_{t-1} - \mu) + \cdots + \phi_p(x_{t-p} - \mu) + w_t + \theta_1 w_{t-1} + \cdots + \theta_q w_{t-q}$$

- Cuando $p = 0$, entonces el modelo ARMA pasa a ser uno de media móvil, mientras que si $q = 0$, el modelo ARMA pasa a ser uno autorregresivo
 - Para poder estudiar de manera más sencilla los modelos ARMA, es necesario utilizar los operadores autorregresivos y de media móvil, expresando el modelo de la siguiente manera:

$$\phi(B)x_t = \theta(B)w_t$$

- Esta representación concisa del modelo permite ver un problema en el que se complica el modelo de más al multiplicar ambos lados por un operador cualquiera $\eta(B)$ sin cambiar las dinámicas

$$\eta(B)\phi(B)x_t = \eta(B)\theta(B)w_t$$

- Si se considera un ruido blanco $x_t = w_t$, se podría multiplicar cada lado de la ecuación por $\eta(B)$ y eso haría que se obtuvieran unos parámetros para los retrasos de cada una de las variables. Estos parámetros, no obstante, enmascaran el proceso simple en uno más complejo, creando sobreparametrización
 - Es posible ajustar un modelo con más parámetros de los necesarios y que estos salgan significativos, de modo que, si no se considerara este problema, se podría decir que existe autocorrelación, aunque no haya
 - Como se puede ver, el uso de modelos AR, MA y ARMA conlleva problemas como la redundancia de parámetros, la unicidad de los modelos MA y la dependencia del futuro de modelos AR explosivos. Para poder superar estos problemas, es necesario poner restricciones adicionales sobre los parámetros del modelo
 - Para poder superar el problema de la redundancia de parámetros, es necesario hacer que el modelo $ARMA(p, q)$ esté en su forma más simple. Esto es equivalente a decir que los operadores $\phi(z)$ y $\theta(z)$ no tienen coeficientes comunes
 - Para poder superar el problema de la dependencia del futuro en los modelos, se introduce el concepto de causalidad. Un modelo $ARMA(p, q)$ es causal si la serie temporal $\{x_t; t = 0, \pm 1, \pm 2, \dots\}$ se puede escribir como el siguiente proceso lineal, llamado $MA(\infty)$:
- $$x_t = \sum_{j=0}^{\infty} \psi_j w_{t-j} = \psi(B)w_t$$
- where* $\psi_0 = 1$ & $\psi(B) = \sum_{j=0}^{\infty} \psi_j B^j$ & $\sum_{j=0}^{\infty} |\psi_j| < \infty$
- Un modelo $ARMA(p, q)$ es causal si, y solo si, $\phi(z) \neq 0$ para $|z| \leq 1$, lo cual es equivalente a decir que es causal cuando

$\phi(z) = 0$ para $|z| > 1$. Los coeficientes del proceso lineal se pueden determinar resolviendo la siguiente ecuación:

$$\psi(z) = \sum_{j=0}^{\infty} \psi_j z^j = \frac{\theta(z)}{\phi(z)} \quad \text{if } |z| < 1 \Leftrightarrow \phi(z) = 0 \text{ if } |z| > 1$$

- Para poder solucionar el problema de la unicidad de los modelos MA, es necesario definir el concepto de invertibilidad. Un modelo $ARMA(p, q)$ es invertible si la serie temporal $\{x_t; t = 0, \pm 1, \pm 2, \dots\}$ se puede escribir como el siguiente proceso lineal, llamado $AR(\infty)$:

$$\pi(B)x_t = \sum_{j=0}^{\infty} \pi_j x_{t-j} = w_t$$

$$\text{where } \pi_0 = 1 \text{ & } \pi(B) = \sum_{j=0}^{\infty} \pi_j B^j \text{ & } \sum_{j=0}^{\infty} |\pi_j| < \infty$$

- Un modelo $ARMA(p, q)$ es invertible si, y solo si, $\theta(z) \neq 0$ para $|z| \leq 1$, lo cual es equivalente a decir que es invertible cuando $\theta(z) = 0$ para $|z| > 1$. Los coeficientes del proceso lineal se pueden determinar resolviendo la siguiente ecuación:

$$\pi(z) = \sum_{j=0}^{\infty} \pi_j z^j = \frac{\phi(z)}{\theta(z)} \quad \text{if } |z| \leq 1 \Leftrightarrow \theta(z) = 0 \text{ if } |z| > 1$$

- Como z es un número complejo, se necesita calcular el módulo para poder saber si $|z| \leq 1$ o no (con números reales, esto no es necesario comprobarlo, ya que el valor absoluto basta)
- A partir de las representaciones del modelo ARMA como $AR(\infty)$ y $MA(\infty)$, se pueden obtener las siguientes propiedades y conclusiones:
 - Todos los modelos $AR(p)$ son invertibles, dado que $\sum_{j=0}^{\infty} |\pi_j| < \infty$ para todos ellos porque $\pi_j = \phi_j$ para $j = 1, 2, \dots, p$
 - Todos los modelos $MA(q)$ son causales, dado que $\sum_{j=0}^{\infty} |\psi_j| < \infty$ para todos ellos porque $\psi_j = \theta_j$ para $j = 1, 2, \dots, q$

Los modelos ARIMA: la autocorrelación

- La función de autocorrelación y la función de autocorrelación parcial son muy importantes para la identificación de los modelos de series temporales, de modo que es necesario estudiar las funciones detalladamente

- Para obtener una expresión de la función de autocorrelación para un modelo $MA(q)$, es necesario derivar la función de autocovarianza

- Debido a que x_t es una combinación lineal de términos de ruido blanco, el proceso es estacionario con la siguiente media y autocovarianza:

$$E(x_t) = E \left[\sum_{j=0}^q \theta_j w_{t-j} \right] = \sum_{j=0}^q \theta_j E(w_{t-j}) = 0$$

$$\gamma(h) = Cov \left(\sum_{j=0}^q \theta_j w_{t+h-j}, \sum_{k=0}^q \theta_k w_{t-k} \right) =$$

$$= E \left[\left(\sum_{j=0}^q \theta_j w_{t+h-j} \right) \left(\sum_{k=0}^q \theta_k w_{t-k} \right) \right] =$$

$$= \begin{cases} \sigma_w^2 \sum_{j=0}^{q-h} \theta_j \theta_{j+h} & \text{if } 0 \leq h \leq q \\ 0 & \text{if } h > q \end{cases}$$

- Recordando que $\gamma(h) = \gamma(-h)$, solo es necesario expresar la ACF en términos de $h \geq 0$, y se puede obtener la siguiente expresión

$$\rho(h) = \frac{\gamma(h)}{\gamma(0)} = \begin{cases} \frac{\sum_{j=0}^{q-h} \theta_j \theta_{j+h}}{1 + \sum_{k=1}^q \theta_k^2} & \text{if } 0 \leq h \leq q \\ 0 & \text{if } h > q \end{cases}$$

- Debido a que $\rho(h)$ es nula para toda $h > q$, la ACF permite obtener una buena idea del orden del modelo $MA(q)$
- Para obtener una expresión de la función de autocorrelación para un modelo $AR(p)$, es necesario generalizar la forma recursiva de la función de autocorrelación a través de la autocovarianza
- Considerando un modelo $AR(p)$, es posible generalizar la expresión para $\gamma(h)$. Sabiendo que $E(x_t) = 0$, entonces se puede obtener la siguiente igualdad:

$$\gamma(h) = E(x_{t+h} x_t) = E(x_t x_{t-h}) =$$

$$= \phi_1 E(x_t x_{t-1}) + \phi_2 E(x_t x_{t-2}) + \cdots + \phi_p E(x_t x_{t-p}) =$$

$$= \phi_1 \gamma(h-1) + \phi_2 \gamma(h-2) + \cdots + \phi_p \gamma(h-p) \text{ for } h \geq 0$$

- Por lo tanto, se puede obtener la siguiente forma recursiva para la función de autocorrelación a través de dividir por $\gamma(0)$

$$\rho(h) = \phi_1 \rho(h-1) + \phi_2 \rho(h-2) + \cdots + \phi_p \rho(h-p)$$

- Si z_1, z_2, \dots, z_r son las raíces de $\phi(z)$, cada una con multiplicidad m_1, m_2, \dots, m_r , en donde $m_1 + m_2 + \cdots + m_r = p$, la solución general será la siguiente, donde $P_j(h)$ es un polinomio en h de grado $m_j - 1$:

$$\rho(h) = z_1^{-h} P_1(h) + z_2^{-h} P_2(h) + \cdots + z_r^{-h} P_r(h) \text{ for } h \geq p$$

- Para un modelo causal, todas las raíces yacen fuera del círculo unitario, por lo que $|z_i| > 1$ para $i = 1, 2, \dots, r$, lo cual tiene implicaciones para el comportamiento de las funciones de autocorrelación
 - Si todas las raíces son reales, entonces $\rho(h)$ decae exponencialmente rápido a cero cuando $h \rightarrow \infty$
 - Si algunas de las raíces son complejas, entonces lo serán en pares conjugados y $\rho(h)$ decae de manera sinusoidal a cero cuando $h \rightarrow \infty$
 - Si todas las raíces son complejas, la función de autocorrelación parecerá cíclica. Esto también aplica a los modelos ARMA cuya parte AR tiene raíces complejas
- Para un modelo $ARMA(p, q)$ causal en donde las raíces de $\phi(z)$ yacen fuera del círculo unitario, es posible obtener una expresión para la función de autocorrelación
 - Escribiendo el modelo $ARMA(p, q)$ como una combinación lineal usando el operador $\psi(z)$, es posible obtener la esperanza y la función de autocovarianza

$$E(x_t) = E\left(\sum_{j=0}^{\infty} \psi_j w_{t-j}\right) = \sum_{j=0}^{\infty} \psi_j E(w_{t-j}) = 0$$

$$\gamma(h) = Cov\left(\sum_{j=0}^{\infty} \psi_j w_{t+h-j}, \sum_{k=0}^{\infty} \psi_k w_{t-k}\right) =$$

$$= \sigma_w^2 \sum_{j=0}^{q-h} \psi_j \psi_{j+h} \text{ for } h \geq 0$$

- Es posible obtener una expresión recursiva para $\gamma(h)$ como en los modelos AR(p)

$$\text{Cov}(w_{t+h-j}, x_t) = \text{Cov}\left(w_{t+h-j}, \sum_{k=0}^{\infty} \psi_k w_{t-k}\right) = \psi_{j-h} \sigma_w^2$$

$$\Rightarrow \gamma(h) = \text{Cov}\left(\sum_{j=1}^p \phi_j x_{t+h-j} + \sum_{j=0}^q \theta_j w_{t-j}, x_t\right) =$$

$$= \sum_{j=1}^p \phi_j \gamma(h-j) + \sigma_w^2 \sum_{j=0}^q \theta_j \psi_{j-h} \text{ for } h \geq 0$$

$$\Rightarrow \gamma(h) = \gamma(h-1) + \phi_2 \gamma(h-2) + \cdots + \phi_p \gamma(h-p)$$

$$\text{for } h \geq \max(p, q + 1)$$

- Dividiendo la expresión recursiva por $\gamma(0)$, es posible obtener una expresión recursiva para $\rho(h)$

$$\rho(h) = \rho(h-1) + \phi_2 \rho(h-2) + \cdots + \phi_p \rho(h-p)$$

$$\text{for } h \geq \max(p, q + 1)$$

- Aunque la ACF es muy útil para identificar el orden para un modelo MA, esta no es muy útil para identificar el orden para modelos AR o ARMA, por lo que es necesario introducir el concepto de la función de autocorrelación parcial

- Si X, Y y Z son variables aleatorias, entonces la correlación parcial entre X e Y condicional a Z se obtiene a través de hacer una regresión de X sobre Z para obtener \hat{X} , otra de Y sobre Z para obtener \hat{Y} , y calcular la siguiente expresión:

$$\rho_{X,Y|Z} = \text{Corr}(X - \hat{X}, Y - \hat{Y})$$

- La idea es que $\rho_{X,Y|Z}$ mide la correlación entre X e Y eliminando parcialmente el efecto lineal de Z . Si las variables siguen una distribución normal multivariante, entonces la definición coincide con la correlación condicionada a Z

- Con tal de motivar esta misma idea para series temporales, se utiliza un modelo $AR(1)$

- Calculando la función de autocovarianza para, por ejemplo, $h = 2$, se obtiene un resultado que muestra como la correlación entre x_{t-2} y x no es nula (es dependiente a través de x_{t-1})

$$\begin{aligned}\gamma_x(2) &= Cov(x_{t+2}, x_t) = Cov(x_t, x_{t-2}) = \\ &= Cov(\phi_1 x_{t-2} + w_{t-1}) + w_t, x_{t-2}) = \phi_1^2 \gamma_x(0)\end{aligned}$$

- Si se elimina el efecto lineal parcial de x_{t-1} , entonces se considera la correlación entre $x_t - \phi_1 x_{t-1}$ y $x_{t-2} - \phi_1 x_{t-1}$ (ya que $E(x_t | x_{t-1}) = \phi_1 x_{t-1}$ y, como no se puede calcular para x_{t-2} , se utiliza el de la primera). Por lo tanto, se obtienen los siguientes resultados:

$$Cov(x_t - \phi_1 x_{t-1}, x_{t-2} - \phi_1 x_{t-1}) = Cov(w_t, x_{t-2} - \phi_1 x_{t-1}) = 0$$

- Por lo tanto, se puede ver como lo que es necesario es la correlación parcial, que es la correlación entre x_t y x_s cuando se elimina el efecto lineal de todos los retrasos del medio

- La función de autocorrelación parcial (PACF) de un proceso estacionario se define para dos casos de la siguiente manera:

$$\phi_{11} = Corr(x_{t+1}, x_t) = \rho(1)$$

$$\phi_{hh} = Corr(x_{t+h} - \hat{x}_{t+h}, x_t - \hat{x}_t) \text{ when } h \geq 2$$

- Para $h \geq 2$, se define \hat{x}_{t+h} como la regresión de x_{t+h} sobre $x_{t+h-1}, x_{t+h-2}, \dots, x_{t+1}$ y se define \hat{x}_{t+h} como la regresión de x_{t+h} sobre $x_{t+1}, x_{t+2}, \dots, x_{t+h-1}$. Los coeficientes de ambas regresiones son los mismos debido a la suposición de estacionariedad de los datos

$$\hat{x}_{t+h} = \beta_1 x_{t+h-1} + \beta_1 x_{t+h-2} + \dots + \beta_1 x_{t+1}$$

$$\hat{x}_t = \beta_1 x_{t+1} + \beta_1 x_{t+2} + \dots + \beta_1 x_{t+h-1}$$

- Si el proceso es normal, entonces la correlación parcial ϕ_{hh} es equivalente a $Corr(x_{t+h}, x_t | x_{t+1}, x_{t+2}, \dots, x_{t+h-1})$, de modo que ϕ_{hh} es el coeficiente de correlación entre x_{t+h} y x_t en la distribución bivariante normal (x_{t+h}, x_t) condicionada a $x_{t+1}, x_{t+2}, \dots, x_{t+h-1}$

- Debido a que un modelo $AR(p)$ causal tiene una representación como una suma finita, es posible encontrar la función de correlación parcial a través de esta

- El modelo implica que $x_{t+h} = \sum_{j=1}^p \phi_j x_{t+h-j} + w_{t+h}$, en donde las raíces están fuera del círculo unitario, de modo que se puede plantear una regresión para x_{t+h} con $h > p$ a partir de la siguiente fórmula (que se demostrará en siguientes secciones):

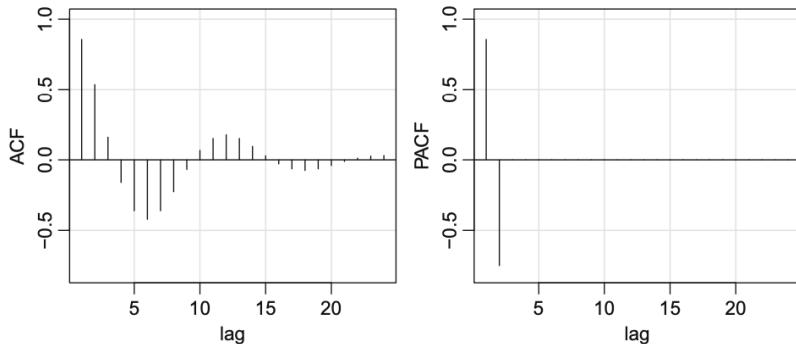
$$\hat{x}_{t+h} = \sum_{j=1}^p \phi_j x_{t+h-j}$$

- Por lo tanto, se puede obtener una correlación nula, dado que, por causalidad, $x_t - \hat{x}_t$ solo depende de $w_{t+h-1}, w_{t+h-2}, \dots$

$$Corr(x_{t+h} - \hat{x}_{t+h}, x_t - \hat{x}_t) = Corr(w_{t+h}, x_t - \hat{x}_t) = 0$$

$$\text{where } \hat{x}_t = \sum_{j=0}^{\infty} \phi_j w_{t+h-j}$$

- Cuando $h \leq p$, $\phi_{pp} \neq 0$ y $\phi_{11}, \phi_{22}, \dots, \phi_{p-1,p-1}$ no son necesariamente nulos

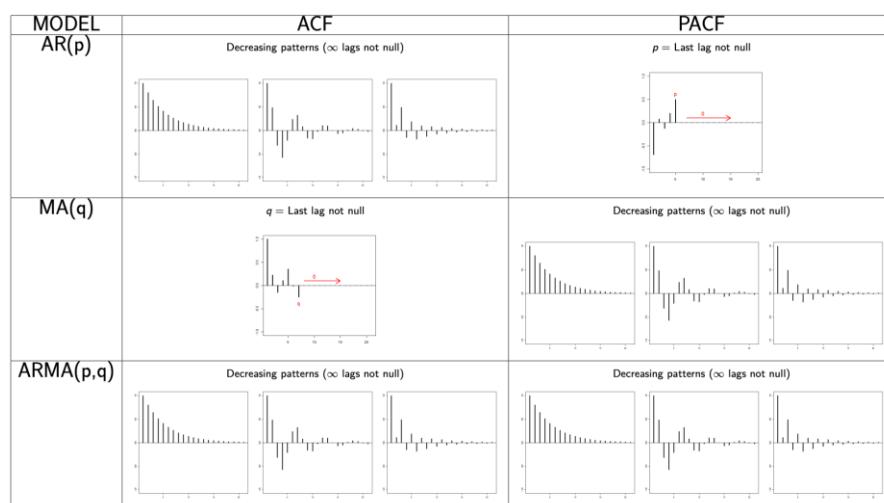


- Debido a que un modelo $MA(q)$ invertible solo tiene una representación como una suma infinita (no existe una representación finita), no es posible encontrar la función de correlación parcial a través de esta

- El modelo implica que $x_{t+h} = -\sum_{j=1}^{\infty} \pi_j x_{t-j} + w_t$, por lo que la PACF general para un modelo $MA(q)$ no se puede calcular una expresión analítica general (dependerá del orden)
- A partir de un modelo $MA(1)$ es posible demostrar que la forma funcional de la PACF es la siguiente:

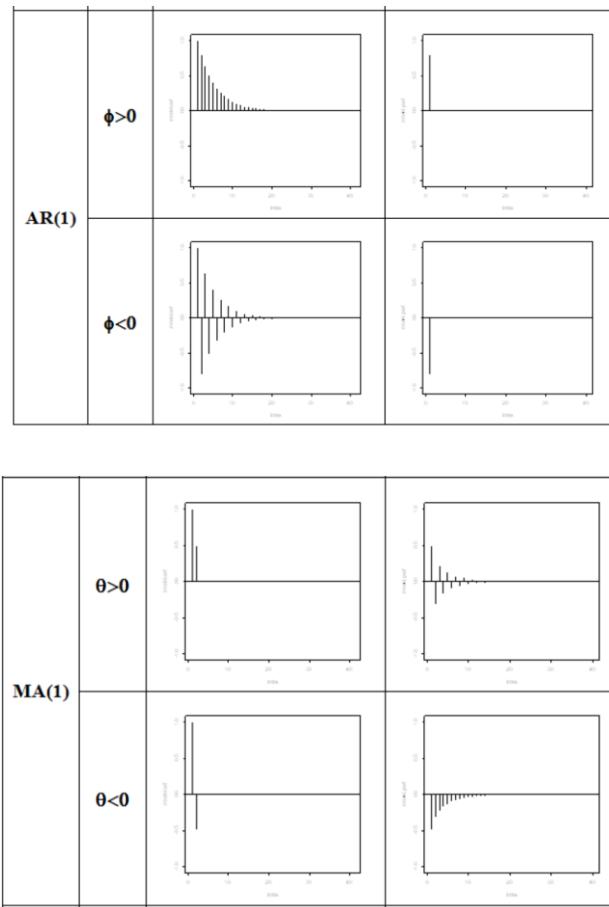
$$\phi_{hh} = \frac{(-\theta)^h(1 - \theta^2)}{1 - \theta^{2(h+1)}} \quad \text{for } h \geq 1$$

- En este caso, la PACF se comportará como una ACF para los modelos AR, de modo que la autocorrelación parcial tiene esas tres dinámicas definidas anteriormente (a partir de ver las raíces)
- La PACF para los modelos MA se comporta como la ACF para los modelos AR, mientras que la PACF para los modelos AR se comporta como la ACF de los modelos MA, por lo que se puede resumir el comportamiento de las funciones en la siguiente tabla:



	AR(p)	MA(q)	ARMA(p, q)
ACF	Tails off	Cuts off after lag q	Tails off
PACF	Cuts off after lag p	Tails off	Tails off

- Debido a que un modelo ARMA invertible tiene una representación AR infinita, la PACF nunca se cortará en un retraso
- La PACF permitirá identificar el orden de un modelo AR, mientras que la ACF permitirá indicar la del modelo MA
- El signo de los valores de la autocorrelación y la autocorrelación parcial dependerán de los signos de los parámetros, de modo que se pueden dar comportamientos mixtos en un ARMA según los parámetros de la parte AR y de la parte MA



Los modelos ARIMA: la estimación

- Existen varias maneras de poder estimar los modelos ARMA vistos hasta ahora basados en métodos estadísticos tradicionales
 - En todo el desarrollo se asume que se tienen n observaciones x_1, x_2, \dots, x_n y se está con un modelo $ARMA(p, q)$ gaussiano causal e invertible en el que se saben los órdenes p y q
 - El objetivo es estimar los parámetros $\phi_1, \phi_2, \dots, \phi_p, \theta_1, \theta_2, \dots, \theta_q$ y σ_w^2
 - El problema de determinar p y q se ha discutido anteriormente (en el contexto de la ACF y la PACF) y se discutirá más adelante
 - Una manera para poder estimar este tipo de modelos es a través del método de momentos, en donde se igualan los momentos poblacionales con los muestrales
 - Debido a que $E(x_t) = \mu = \bar{x}$, entonces se asume que $\mu = 0$

- Aunque este método permite obtener buenos estimadores, el método puede llevar a considerar estimadores subóptimos
- Para modelos $AR(p)$, los estimadores por método de momentos serán eficientes. Para estimarlos, se utilizan las ecuaciones de Yule-Walker, las cuales son las siguientes:

$$\gamma(h) = \phi_1\gamma(h-1) + \cdots + \phi_p\gamma(h-p) \quad \text{for } h = 1, 2, \dots, p$$

$$\sigma_w^2 = \gamma(0) - \phi_1\gamma(1) - \cdots - \phi_p\gamma(p)$$

- En notación matricial, las ecuaciones de Yule-Walker se pueden expresar a través de una matriz $\Gamma_p = \{\gamma(k-j)\}_{j,k=1}^p$ de tamaño $p \times p$, un vector $\boldsymbol{\phi} = (\phi_1, \phi_2, \dots, \phi_p)'$ de tamaño $p \times 1$ y un vector $\boldsymbol{\gamma}_p = (\gamma(0), \gamma(1), \dots, \gamma(p))'$ de tamaño $p \times 1$

$$\boldsymbol{\Gamma}_p \boldsymbol{\phi} = \boldsymbol{\gamma}_p \quad \sigma_w^2 = \gamma(0) - \boldsymbol{\gamma}_p' \boldsymbol{\phi}$$

- Estas ecuaciones nacen de las ecuaciones generales homogéneas que se pueden derivar para la varianza σ_w^2 y para la función de autocovarianza para un proceso ARMA (sus formas recursivas)
- Usando el método de momentos, se reemplaza $\gamma(h)$ por $\hat{\gamma}(h)$ y se resuelve para los parámetros deseados, obteniendo los estimadores de Yule-Walker

$$\hat{\boldsymbol{\phi}} = \hat{\boldsymbol{\Gamma}}_p^{-1} \hat{\boldsymbol{\gamma}}_p \quad \hat{\sigma}_w^2 = \hat{\gamma}(0) - \hat{\boldsymbol{\gamma}}_p' \hat{\boldsymbol{\Gamma}}_p^{-1} \hat{\boldsymbol{\gamma}}_p$$

- Por razones de cálculo, es más conveniente trabajar con la función de autocorrelación parcial. Al factorizar $\hat{\gamma}(0)$ en las ecuaciones anteriores, se pueden escribir los estimadores de Yule-Walker a través de una matriz $\hat{\boldsymbol{R}}_p = \{\hat{\rho}(k-j)\}_{j,k=1}^p$ de tamaño $p \times p$ y un vector $\hat{\boldsymbol{\rho}}_p = (\hat{\rho}(1), \hat{\rho}(2), \dots, \hat{\rho}(p))'$ de tamaño $p \times 1$

$$\hat{\boldsymbol{\phi}} = \hat{\boldsymbol{R}}_p^{-1} \hat{\boldsymbol{\rho}}_p \quad \hat{\sigma}_w^2 = \hat{\gamma}(0)[1 - \hat{\boldsymbol{\rho}}_p' \hat{\boldsymbol{R}}_p^{-1} \hat{\boldsymbol{\rho}}_p]$$

- Para modelos $AR(p)$, si el tamaño muestral es grande, los estimadores de Yule-Walker se distribuyen aproximadamente normal y $\hat{\sigma}_w^2$ está cerca del verdadero valor de c
 - El comportamiento asintótico de los estimadores de Yule-Walker en el caso de un modelo $AR(p)$ causal es el siguiente:

$$\sqrt{n}(\hat{\boldsymbol{\phi}} - \boldsymbol{\phi}) \xrightarrow{d} N_p(0, \sigma_w^2 \boldsymbol{\Gamma}_p^{-1}) \quad \hat{\sigma}_w^2 \xrightarrow{p} \sigma_w^2$$

- Este comportamiento asintótico muestra como los estimadores de Yule-Walker son eficientes, dado que, en verdad, un modelo $AR(p)$ causal es un modelo lineal y los estimadores de Yule-Walker son como estimadores de mínimos cuadrados
- El algoritmo de Durbin-Levinson se puede usar para calcular $\hat{\boldsymbol{\phi}}$ sin invertir $\boldsymbol{\Gamma}_p^{-1}$ o $\hat{\boldsymbol{R}}_p^{-1}$, al reemplazar $\gamma(h)$ y $\hat{\gamma}(h)$ en el algoritmo
 - En este algoritmo, se calcula iterativamente el vector $\hat{\boldsymbol{\phi}}_h = (\hat{\phi}_{h1}, \hat{\phi}_{h2}, \dots, \hat{\phi}_{hh})'$ de tamaño $h \times 1$ para $h = 1, 2, \dots$, por lo que se pueden obtener las predicciones y la PACF muestral
 - Para un modelo $AR(p)$ causal, el comportamiento asintótico de la PACF muestral es el siguiente:
- $\sqrt{n}\hat{\phi}_{hh} \xrightarrow{d} N(0,1) \text{ for } h > p$
- Para modelos $MA(q)$ y $ARMA(p,q)$, los estimadores del método de momentos no serán óptimos porque los procesos no son lineales en sus parámetros. Se puede ejemplificar esto con un modelo $MA(1)$
- Otra manera para poder estimar modelos es través de método de máxima verosimilitud
 - Para fijar ideas, se ejemplifica todo a partir de un modelo $AR(1)$ como $x_t = \mu + \phi_1(x_{t-1} - \mu) + w_t$ en donde $|\phi_1| < 1$ y $w_t \sim iid N(0, \sigma_w^2)$. Dados los datos x_1, x_2, \dots, x_n , se busca una función de verosimilitud como la siguiente:

$$L(\mu, \phi_1, \sigma_w^2) = f(x_1, x_2, \dots, x_n | \mu, \phi_1, \sigma_w^2)$$
 - En este caso de un $AR(1)$, se puede escribir la función de verosimilitud, eliminando los parámetros en las densidades con tal de simplificar la notación:
$$L(\mu, \phi_1, \sigma_w^2) = f(x_1)f(x_2|x_1) \dots f(x_n|x_{n-1})$$
 - Como $x_t|x_{t-1} \sim N(\mu + \phi_1(x_{t-1} - \mu), \sigma_w^2)$, la función de verosimilitud se puede expresar en términos de la función de densidad del ruido blanco $f_w(\cdot)$ a través de $x_t|x_{t-1}$
$$f(x_n|x_{n-1}) = f_w[(x_t - \mu) - \phi_1(x_{t-1} - \mu)]$$

$$\Rightarrow L(\mu, \phi_1, \sigma_w^2) = f(x_1) \prod_{i=2}^n f_w[(x_t - \mu) - \phi_1(x_{t-1} - \mu)]$$

- Bajo condiciones apropiadas para modelos causales e invertibles ARMA, los estimadores de máxima verosimilitud y los mínimos cuadrados condicionales e incondicionales, inicializándose por el estimador de método de momentos, proporcionan estimadores óptimos de σ_w^2 y β

$$\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{d} N_{p+q}(0, \sigma_w^2 \Gamma_{p,q}^{-1})$$

- Los estimadores son óptimos en el sentido de que $\hat{\sigma}_w^2$ es consistente (tiende en probabilidad a σ_w^2) y la distribución asintótica de $\hat{\beta}$ es normal multivariante con media β y matriz de varianzas y covarianzas $\sigma_w^2 \Gamma_{p,q}^{-1}$
- La matriz de varianzas y covarianzas asintótica del estimador $\hat{\beta}$ es la inversa de la matriz de información $E\left[-\frac{\partial^2}{\partial \beta^2} \log(L(\beta, \sigma_w^2))\right]$. En particular, la matriz $\Gamma_{p,q}$ de tamaño $(p+q) \times (p+q)$ tiene la siguiente forma:

$$\Gamma_{p,q} = \begin{pmatrix} \Gamma_{\phi\phi} & \Gamma_{\phi\theta} \\ \Gamma_{\theta\phi} & \Gamma_{\theta\theta} \end{pmatrix}$$

- Esta matriz está compuesta por las matrices $\Gamma_{\phi\phi} = \{\gamma_x(i-j)\}_{j,k=1}^p$, $\Gamma_{\theta\theta} = \{\gamma_y(i-j)\}_{i,k=1}^q$, $\Gamma_{\phi\theta} = \{\gamma_{xy}(i-j)\}$ para $i = 1, 2, \dots, p$ y $j = 1, 2, \dots, q$ (de tamaño $p \times q$) y $\Gamma_{\theta\phi} = \Gamma'_{\phi\theta}$

Los modelos ARIMA: la predicción

- El objetivo de la predicción es predecir los futuros x_{n+m} para $m = 1, 2, \dots$ de una serie temporal basándose en los valores presentes $\mathbf{x}_{1:n} = \{x_1, x_2, \dots, x_n\}$. Esto se desarrolla a través de asumir que x_t es estacionaria y los parámetros se conocen
 - El predictor de error de predicción mínimo de x_{n+m} se define de la siguiente manera:

$$x_{n+m}^n = E(x_{n+m} | \mathbf{x}_{1:n}) \text{ for } m = 1, 2, \dots$$

- Esta esperanza condicional es el valor que minimiza el error cuadrático medio de predicción, en donde $g(\mathbf{x}_{1:n})$ es una función de las observaciones $\{x_1, x_2, \dots, x_n\}$

$$E\left[\left(x_{n+m} - g(\mathbf{x}_{1:n})\right)^2\right]$$

- Primero se suele restringir la atención a predictores que son funciones lineales de los datos, en donde los parámetros $\alpha_0, \alpha_1, \alpha_2, \dots, \alpha_n$ son números reales

$$x_{n+m}^n = \alpha_0 + \sum_{k=1}^n \alpha_k x_k$$

- Los predictores lineales solo dependen de los segundos momentos del proceso, los cuales son fáciles de estimar a partir de los datos
- Estos parámetros dependen de n y m , pero no se representa la dependencia de manera explícita. Por ejemplo, si $n = m = 1$, entonces $x_2^1 = \alpha_0 + \alpha_1 x_1$, pero si $n = 2$, entonces $x_3^2 = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2$, en donde las α generalmente serán diferentes (dado que ahora se tiene en cuenta el efecto de x adicionales)
- Los predictores lineales de la forma mostrada que minimizan el error de predicción medio cuadrático se denominan mejores predictores lineales o *best linear predictors* (BLP)
- Dados los datos x_1, x_2, \dots, x_n , el mejor predictor lineal $x_{n+m}^n = \alpha_0 + \sum_{k=1}^n \alpha_k x_k$ de x_{n+m} para $m \geq 1$ se encuentra resolviendo la siguiente ecuación para $\alpha_0, \alpha_1, \alpha_2, \dots, \alpha_n$

$$E[(x_{n+m} - x_{n+m}^n)x_k] = 0 \quad \text{for } k = 0, 1, \dots, n \quad \& \quad x_0 = 1$$
 - Este resultado se apoya en el teorema de la proyección de la teoría del dominio temporal
 - El teorema muestra un sistema de ecuaciones con las llamadas ecuaciones de predicción, con las que se pueden obtener los coeficientes $\alpha_0, \alpha_1, \alpha_2, \dots, \alpha_n$
 - Los resultados de esta propiedad también se pueden obtener a través de mínimos cuadrados minimizando $E[(x_{n+m} - \sum_{k=1}^n \alpha_k x_k)^2]$ con respecto a las α
 - Si $E(x_t) = \mu$, la primera ecuación $k = 0$ permite obtener la siguiente forma para el mejor predictor lineal:

$$E[(x_{n+m} - x_{n+m}^n)x_0] = 0 \Rightarrow E(x_{n+m} - x_{n+m}^n) = 0$$

$$\Rightarrow E(x_{n+m}) - E(x_{n+m}^n) = 0 \Rightarrow E(x_{n+m}) = E(x_{n+m}^n) = \mu$$

$$E(x_{n+m}^n) = \alpha_0 + \sum_{k=1}^n \alpha_k \mu \Rightarrow \mu = \alpha_0 + \sum_{k=1}^n \alpha_k \mu$$

$$\Rightarrow \alpha_0 = \mu \left(1 - \sum_{k=1}^n \alpha_k \right)$$

$$\Rightarrow x_{n+m}^n = \mu \left(1 - \sum_{k=1}^n \alpha_k \right) + \sum_{k=1}^n \alpha_k x_k = \mu + \sum_{k=1}^n \alpha_k (x_k - \mu)$$

- Gracias al resultado anterior, no se pierde generalidad si se asume que $\mu = 0$ para el desarrollo teórico siguiente
- Considerando la predicción a un paso adelante (predecir el valor x_{n+1}), el mejor predictor lineal de x_{n+1} tiene la siguiente forma:

$$x_{n+1}^n = \phi_{n1} x_n + \phi_{n2} x_{n-1} + \cdots + \phi_{nn} x_1$$

- Ahora si se muestra la dependencia de los coeficientes en n y α_k para a ser $\phi_{n,n+1-k}$ para $k = 1, 2, \dots, n$
- Usando lo propuesto anterior, los coeficientes $\{\phi_{n1}, \phi_{n2}, \dots, \phi_{nn}\}$ satisfacen las siguientes ecuaciones:

$$E \left[\left(x_{n+1} - \sum_{j=1}^n \phi_{nj} x_{n+1-j} \right) x_{n+1-k} \right] = 0 \quad \text{for } k = 1, 2, \dots, n$$

$$\sum_{j=1}^n \phi_{nj} \gamma(k-j) = \gamma(k) \quad \text{for } k = 1, 2, \dots, n$$

- Las ecuaciones de predicción anteriores se pueden escribir con notación matricial a partir del segundo conjunto de ecuaciones, en donde $\Gamma_n = \{\gamma(k-j)\}_{j,k=1}^n$ es una matriz semidefinida positiva de tamaño $n \times n$, $\boldsymbol{\phi}_n = (\phi_{n1}, \phi_{n2}, \dots, \phi_{nn})'$ es un vector $n \times 1$ y $\boldsymbol{\gamma}_n = (\gamma(1), \gamma(2), \dots, \gamma(n))'$ es un vector de tamaño $n \times 1$

$$\boldsymbol{\Gamma}_n \boldsymbol{\phi}_n = \boldsymbol{\gamma}_n$$

- Si $\boldsymbol{\Gamma}_n$ no es invertible, entonces existen varias soluciones para $\boldsymbol{\Gamma}_n \boldsymbol{\phi}_n = \boldsymbol{\gamma}_n$, pero, por el teorema de proyección, x_{n+1}^n es único. Si, en cambio, $\boldsymbol{\Gamma}_n$ es invertible, los elementos de $\boldsymbol{\phi}_n$ son únicos y se dan por la siguiente ecuación:

$$\phi_n = \Gamma_n^{-1} \gamma_n$$

- Para los modelos ARMA, el hecho de que $\sigma_w^2 > 0$ y $\gamma(h) \rightarrow 0$ cuando $h \rightarrow \infty$ es suficiente para asegurar que Γ_n es definida positiva
- A veces es conveniente escribir la predicción de un paso adelante en forma vectorial

$$x_{n+1}^n = \phi_n' x \text{ where } x = (x_1, x_2, \dots, x_n)$$

- El error de predicción cuadrático medio para la predicción un paso adelante es la siguiente:

$$P_{n+1}^n = E[(x_{n+1} - x_{n+1}^n)^2] = \gamma(0) - \gamma_n' \Gamma_n^{-1} \gamma_n$$

- La demostración de este resultado es la siguiente:

$$\begin{aligned} E[(x_{n+1} - x_{n+1}^n)^2] &= E[(x_{n+1} - \phi_n' x)^2] = \\ &= E[(x_{n+1} - \gamma_n' \Gamma_n^{-1} x)^2] = E[x_{n+1}^2 - 2x_{n+1} \gamma_n' \Gamma_n^{-1} x x' \Gamma_n^{-1} \gamma_n] = \\ &= E[x_{n+1}^2 - 2x_{n+1} \gamma_n' \Gamma_n^{-1} x + \gamma_n' \Gamma_n^{-1} x x' \Gamma_n^{-1} \gamma_n] \\ &= \gamma(0) - 2\gamma_n' \Gamma_n^{-1} \gamma_n + \gamma_n' \Gamma_n^{-1} \Gamma_n \Gamma_n^{-1} \gamma_n \\ &= \gamma(0) - \gamma_n' \Gamma_n^{-1} \gamma_n \end{aligned}$$

- Las ecuaciones de predicción generales no proporcionan una buena perspectiva sobre cómo predecir valores con modelos ARMA generales, de modo que se desarrolla como hacer predicción a partir de los modelos ARMA

- Hay varias maneras diferentes de expresar estas predicciones, y cada una ayuda a entender la estructura especial de la predicción con modelos ARMA
 - En todo el desarrollo se asume que x_t es un proceso ARMA(p, q) causal e invertible

$$\phi(B)x_t = \theta(B)w_t \text{ where } w_t \sim iid N(0, \sigma_w^2)$$

- En el caso en el que $E(x_t) = \mu_x$, se reemplaza x_t por $x_t - \mu_x$ en el modelo
- Para poder analizar la predicción con modelos ARMA, primero se consideran dos tipos de predicciones y se define la notación que se va a utilizar

- Se define x_{n+m}^n como el predictor de error medio cuadrático mínimo de x_{n+m} basado en los datos $x_n \dots, x_1$, expresado de la siguiente manera:

$$x_{n+m}^n \equiv E(x_{n+m}|x_n \dots, x_1)$$

- Para procesos ARMA es más fácil calcular el predictor de x_{n+m} asumiendo que se tiene la historia completa del proceso $\{x_n, x_{n-1}, \dots, x_1, x_0, x_{-1}, \dots\}$. Se denota el predictor de x_{n+m} basado en su pasado infinito de la siguiente manera:

$$\tilde{x}_{n+m} \equiv E(x_{n+m}|x_n, x_{n-1}, \dots, x_1, x_0, x_{-1}, \dots)$$

- En general x_{n+m}^n y \tilde{x}_{n+m} no son lo mismo, pero para muestras grandes se puede ver como x_{n+m}^n proporciona una buena aproximación para \tilde{x}_{n+m}

- Tomando la esperanza condicional sobre la historia completa de la serie, se pueden obtener los siguientes resultados:

- Tomando la esperanza condicional sobre la historia completa de la serie en la forma causal del modelo ARMA, se puede representar la forma causal del modelo (gracias a las propiedades de causalidad e invertibilidad) de la siguiente manera:

$$x_{n+m} = \sum_{j=0}^{\infty} \psi_j w_{n+m-j} \text{ where } \psi_0 = 1$$

$$\Rightarrow \tilde{x}_{n+m} = \sum_{j=0}^{\infty} \psi_j \tilde{w}_{n+m-j} = \sum_{j=m}^{\infty} \psi_j w_{n+m-j} =$$

$$= \tilde{w}_{n+m} + \psi_1 \tilde{w}_{n+m-1} + \dots + \psi_{m-1} \tilde{w}_{n+1} + \psi_m \tilde{w}_n + \dots$$

$$\Rightarrow \tilde{x}_{n+m} = \sum_{j=m}^{\infty} \psi_j w_{n+m-j} = \psi_m w_n + \psi_{m+1} w_{n-1} + \dots$$

$$\text{as } \tilde{w}_t = E(w_t|x_n, x_{n-1}, \dots, x_1, x_0, x_{-1}, \dots) = \begin{cases} 0 & t > n \\ w_t & t \leq n \end{cases}$$

- Tomando la esperanza condicional sobre la historia completa de la serie en la forma invertible del modelo ARMA, se puede representar la forma invertible del modelo (gracias a las

propiedades de causalidad e invertibilidad) de la siguiente manera:

$$\begin{aligned}
 w_{n+m} &= \sum_{j=0}^{\infty} \pi_j x_{n+m-j} \quad \text{where } \pi_0 = 1 \\
 \Rightarrow \tilde{w}_{n+m} &= \sum_{j=0}^{\infty} \pi_j \tilde{x}_{n+m-j} = \tilde{x}_{n+m} + \sum_{j=1}^{\infty} \pi_j \tilde{x}_{n+m-j} = 0 \\
 \Rightarrow \tilde{x}_{n+m} &= - \sum_{j=1}^{m-1} \pi_j \tilde{x}_{n+m-j} - \sum_{j=m}^{\infty} \pi_j x_{n+m-j} = \\
 &= -\pi_1 \tilde{x}_{n+m-1} - \cdots - \pi_{m-1} \tilde{x}_{n+1} - \pi_m x_n - \pi_{m+1} x_{n-1} \dots
 \end{aligned}$$

$$\text{as } \tilde{x}_t = E(x_t | x_n, x_{n-1}, \dots, x_1, x_0, x_{-1}, \dots) = x_t \text{ for } t \leq n$$

- La representación $MA(\infty)$ es muy útil debido a que se puede truncar para valores pequeños de ψ_j y permite calcular de manera sencilla la varianza de la predicción. Usando el resultado obtenido con la forma causal, se puede obtener una identidad para $x_{n+m} - \tilde{x}_{n+m}$ que permite obtener el error de predicción medio cuadrático

$$\begin{aligned}
 x_{n+m} - \tilde{x}_{n+m} &= \sum_{j=0}^{\infty} \psi_j w_{n+m-j} - \sum_{j=m}^{\infty} \psi_j w_{n+m-j} = \\
 &= \sum_{j=0}^{m-1} \psi_j w_{n+m-j} = w_{n+m} + \psi_1 w_{n+m-1} + \cdots + \psi_{m-1} w_{n+1} \\
 \Rightarrow P_{n+m}^n &= E[(x_{n+m} - \tilde{x}_{n+m})^2] = \sigma_w^2 \sum_{j=0}^{m-1} \psi_j^2 \quad \text{where } \psi_0 = 1
 \end{aligned}$$

- Para una muestra fija n , los errores de predicción están correlacionados. Cuando $k \geq 1$, se puede calcular la autocovarianza de los errores de predicción de la siguiente forma:

$$E[(x_{n+m} - \tilde{x}_{n+m})(x_{n+m+k} - \tilde{x}_{n+m+k})] = \sigma_w^2 \sum_{j=0}^{m-1} \psi_j \psi_{j+k}$$

- A veces también es útil representar el modelo $ARMA(p, q)$ en su forma estándar para poder operar con este

$$\phi(B)x_t = \theta(B)w_t \Rightarrow \phi(B)x_{n+m} = \theta(B)w_{n+m}$$

- Esta forma se puede aplicar de manera iterativa con tal de obtener las predicciones (a través de aplicar la esperanza condicional)

$$\tilde{x}_{n+1} = \phi_1 x_n + \phi_2 x_{n-1} + \cdots + \theta_1 \tilde{w}_{n+1} + \theta_2 w_n + \cdots$$

$$\tilde{x}_{n+2} = \phi_1 \tilde{x}_{n+1} + \phi_2 x_n + \cdots + \theta_1 \tilde{w}_{n+2} + \theta_2 \tilde{w}_{n+1} + \cdots$$

...

$$\phi(B)\tilde{x}_{n+m} = \theta(B)\tilde{w}_{n+m}$$

- Considerando la predicción para un proceso ARMA con media μ_x , reemplazando x_{n+m} por $x_{n+m} - \mu_x$ y tomando esperanzas condicionales a la historia completa del proceso, se puede deducir que la predicción para el paso m se puede escribir de la siguiente manera:

$$\tilde{x}_{n+m} = \mu_x + \sum_{j=m}^{\infty} \psi_j w_{n+m-j}$$

- Sabiendo que las ponderaciones ψ tienden a cero de manera exponencialmente rápida, se puede ver como $\tilde{x}_{n+m} \rightarrow \mu_x$ cuando $m \rightarrow \infty$
- Además, el error medio cuadrático se comporta asintóticamente de la siguiente manera cuando $m \rightarrow \infty$

$$P_{n+m}^n \rightarrow \sigma_w^2 \sum_{j=0}^{m-1} \psi_j^2 = \gamma_x(0) = \sigma_x^2$$

- Por lo tanto, las predicciones de un proceso ARMA tienden rápido a la media con un error de predicción constante cuando $m \rightarrow \infty$
- Cuando n es pequeña, las ecuaciones de predicción generales pueden usarse fácilmente, pero cuando n es grande, se utilizaría los resultados obtenidos con la forma causal del proceso, ya que no se observan datos pasados a $x_0, x_{-1}, x_{-2}, \dots$ y solo hay disponibles x_1, x_2, \dots, x_n
- En este caso, se puede truncar $\tilde{x}_{n+m} = -\sum_{j=1}^{m-1} \pi_j \tilde{x}_{n+m-j} - \sum_{j=m}^{\infty} \pi_j x_{n+m-j}$ fijando $\sum_{j=n+m}^{\infty} \pi_j x_{n+m-j} = 0$, por lo que la

representación $AR(\infty)$ permite calcular las predicciones puntuales fácilmente truncando para valores pequeños de π_j . Por lo tanto, el predictor truncado se escribe de la siguiente manera:

$$\tilde{x}_{n+m}^n = - \sum_{j=1}^{m-1} \pi_j \tilde{x}_{n+m-j}^n - \sum_{j=m}^{n+m-1} \pi_j x_{n+m-j}$$

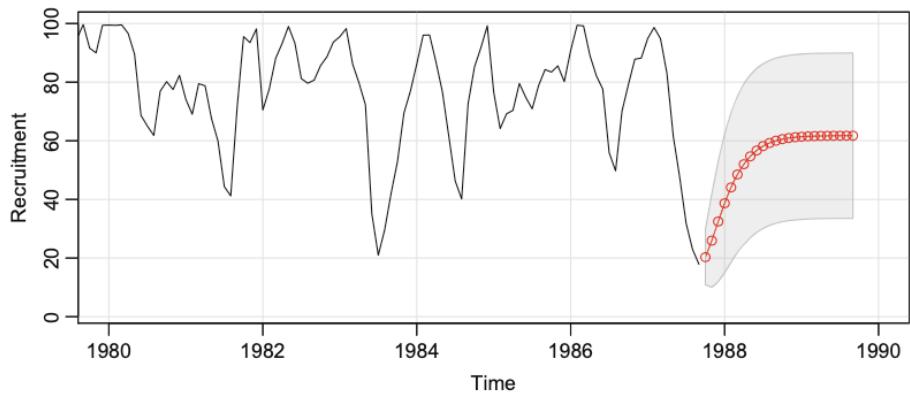
- Para modelos $ARMA(p, q)$, el predictor truncado para $m = 1, 2, \dots$ son los siguientes, en donde $\tilde{x}_t^n = x_t$ para $1 \leq t \leq n$ y $\tilde{x}_t^n = 0$ para $t \leq 0$:

$$\tilde{x}_{n+m}^n = \phi_1 \tilde{x}_{n+m-1}^n + \cdots + \phi_p \tilde{x}_{n+m-p}^n + \theta_1 \tilde{w}_{n+m-1}^n + \cdots + \theta_q \tilde{w}_{n+m-q}^n$$

- Los errores de predicción truncados se dan por $\tilde{w}_t^n = 0$ para $t \leq 0$ o para $t > n$, y por la siguiente expresión para $1 \leq t \leq n$:

$$\tilde{w}_t^n = \phi(B) \tilde{x}_t^n - \theta_1 \tilde{w}_{t-1}^n - \theta_2 \tilde{w}_{t-2}^n - \cdots - \theta_q \tilde{w}_{t-q}^n$$

- Para poder evaluar la precisión de las predicciones, normalmente se utilizan los intervalos de predicción normalmente se calculan junto a las predicciones



- En general, los intervalos de predicción $(1 - \alpha)$ intervalos de predicción son de la siguiente forma:

$$x_{n+m}^n \pm c_{\alpha/2} \sqrt{P_{n+m}^n}$$

- Si se asume que el proceso es gaussiano, normalmente se utiliza un intervalo del 95% de confianza y se fija $c_{\alpha/2} = 1.96$
- Si uno está interesado en establecer intervalos de predicción para más de un periodo de tiempo, entonces $c_{\alpha/2}$ se debe ajustar apropiadamente usando métodos como la desigualdad de Bonferroni

Los modelos ARIMA: los modelos no estacionarios

- En muchas situaciones, las series temporales pueden entenderse como la suma de dos componentes: un componente de tendencia no estacionaria μ_t y un componente estacionario de media nula y_t

$$x_t = \mu_t + y_t$$

- Se pueden plantear varios casos en los que la diferenciación de las series comporta pasar de un proceso no estacionario a uno estacionario
 - Si se considera una serie temporal con tendencia lineal $\mu_t = \beta_0 + \beta_1 t$ e y_t estacionaria, diferenciar este proceso no estacionario una vez permite obtener un proceso estacionario

$$\nabla x_t = x_t - x_{t-1} = \beta_0 + \beta_1 t + y_t - \beta_0 - \beta_1 t + \beta_1 - y_{t-1}$$

$$\Rightarrow \nabla x_t = \beta_1 + \nabla y_t$$

- Si se considera una serie temporal con tendencia de camino aleatorio $\mu_t = \mu_{t-1} + v_t$ y v_t y y_t estacionaria, diferenciar este proceso no estacionario una vez permite obtener un proceso estacionario

$$\nabla x_t = x_t - x_{t-1} = \mu_{t-1} + v_t + y_t - \mu_{t-1} - y_{t-1}$$

$$\Rightarrow \nabla x_t = v_t + \nabla y_t$$

- Si se considera una serie temporal con tendencia polinómica $\mu_t = \sum_{j=0}^k \beta_j t^j$ u otros modelos estocásticos para μ_t , se puede demostrar que las series diferenciadas en un orden mayor ($\nabla^k x_t$ u otro) son estacionarias

$$\mu_t = \mu_{t-1} + v_t \quad \text{where} \quad v_t = v_{t-1} + e_t$$

$$\Rightarrow \nabla x_t = v_t + \nabla y_t = v_{t-1} + e_t + \nabla y_t$$

$$\Rightarrow \nabla^2 x_t = v_{t-1} + e_t + \nabla y_t - v_{t-1} - \nabla y_{t-1} = e_t + \nabla^2 y_t$$

- Por lo tanto, se pueden generalizar los modelos ARMA para tener en cuenta la integración e incluye la diferenciación necesaria para la estacionariedad

- Un proceso x_t es un $ARIMA(p-d, d, q)$ si $\nabla^d x_t = (1-B)^d x_t$ es un $ARMA(p-d, q)$. En general, se escribe el modelo de la siguiente manera:

$$\phi(B)(1 - B)^d x_t = \theta(B)w_t$$

- Si $E(\nabla^d x_t) = \mu$, entonces el modelo se puede escribir de la siguiente manera, donde $\delta = \mu(1 - \phi_1 - \phi_2 - \dots - \phi_p)$:

$$\phi(B)(1 - B)^d x_t = \delta + \theta(B)w_t$$

- Debido a la no estacionariedad, las predicciones se deben hacer de manera más cuidadosa. Como $y_t = \nabla^d x_t$ es un ARMA, se pueden usar los métodos explicados anteriormente para hacer predicciones de y_t , lo cual lleva a obtener predicciones para x_t

- Para $d = 1$, dadas las predicciones y_{n+m}^n para $m = 1, 2, \dots$, se tiene que $y_{n+m}^n = x_{n+m}^n - x_{n+m-1}^n$, de modo que se obtiene la identidad $x_{n+m}^n = y_{n+m}^n + x_{n+m-1}^n$ con condición inicial $x_{n+1}^n = y_{n+1}^n + x_n$

$$x_{n+m}^n = y_{n+m}^n + x_{n+m-1}^n$$

- Para hacer la predicción con esta fórmula, se tiene que obtener primero una predicción para y_{n+1}^n , lo cual se puede hacer de manera recursiva con los datos

$$y_{n+m+k}^n = x_{n+m+k}^n - x_{n+m+k-1}^n$$

- Esta forma se puede aplicar de manera iterativa con tal de obtener las predicciones (a través de aplicar la esperanza condicional)

$$\tilde{x}_{n+m+1} = \phi_1 x_{n+m} + \phi_2 x_{n+m-1} + \dots + \theta_1 \tilde{w}_{n+m+1} + \theta_2 w_{n+m} + \dots$$

$$\tilde{x}_{n+m+2} = \phi_1 \tilde{x}_{n+m+1} + \phi_2 x_{n+m} + \dots + \theta_1 \tilde{w}_{n+m+1} + \theta_2 w_{n+m} + \dots$$

...

$$\phi(B)\tilde{x}_{n+m+k} \dots = \theta(B)\tilde{w}_{n+m+k}$$

- Es más difícil obtener los errores de predicción P_{n+m}^n , pero para una n grande, la aproximación usada anteriormente funciona bien, de modo que el error de predicción cuadrático medio se puede aproximar por la siguiente ecuación:

$$P_{n+m}^n = \sigma_w^2 \sum_{j=0}^{m-1} \psi_j^{*2} \quad \text{where} \quad \psi_j^*(z) = \frac{\theta(z)}{\phi(z)(1-z)^d}$$

- Es posible introducir varias modificaciones para el modelo ARIMA que tienen en cuenta la estacionalidad y el comportamiento no estacionario a través del modelo SARIMA
 - Normalmente, la dependencia en el pasado tiende a ocurrir de manera más pronunciada en múltiplos de algún retraso estacional s
 - Con series temporales económicas mensuales, hay un fuerte componente anual que ocurren en los retrasos múltiples de $s = 12$ debido a las fuertes conexiones de toda actividad con el año del calendario
 - Los datos que se toman de manera trimestral exhiben un periodo anual repetitivo en $s = 4$
 - Por lo tanto, la variabilidad natural de muchos procesos económicos, físicos y biológicos tienden a concordar con fluctuaciones estacionales
 - Esta es la razón por lo que es apropiado introducir polinomios autorregresivos y de media móvil que identifiquen los retrasos estacionales
 - El modelo resultante es el modelo ARMA puramente estacional, denotado por $ARMA(P, Q)$, el cual tiene la siguiente forma:
- $$\Phi_P(B^s)x_t = \Theta_Q(B^s)w_t$$
- Este modelo se utiliza el operador autorregresivo estacional y el operador de media móvil estacional, definidos de la siguiente manera:

$$\Phi_P(B^s) = 1 - \Phi_1 B^s - \Phi_2 B^{2s} - \dots - \Phi_P^{Ps}$$

$$\Theta_Q(B^s) = 1 + \Theta_1 B^s + \Theta_2 B^{2s} + \dots + \Theta_Q B^{Qs}$$
 - De manera análoga a las propiedades de los modelos ARMA no estacionales, el modelo $ARMA(P, Q)$ es causal solo cuando las raíces de $\Phi_P(z^s)$ están fuera del círculo unitario y es invertible solo cuando las raíces de $\Theta_Q(z^s)$ están fuera del círculo unitario
 - Como un criterio de diagnóstico inicial, se pueden usar las propiedades para los modelos autorregresivos y de media móvil puramente estacionales para identificar el orden P y Q

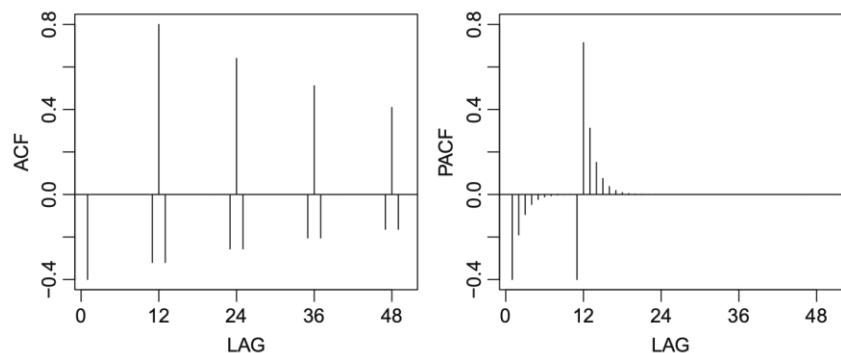
	$AR(P)_s$	$MA(Q)_s$	$ARMA(P, Q)_s$
ACF*	Tails off at lags ks , $k = 1, 2, \dots$,	Cuts off after lag Qs	Tails off at lags ks
PACF*		Tails off at lags ks $k = 1, 2, \dots$,	Tails off at lags ks

*The values at nonseasonal lags $h \neq ks$, for $k = 1, 2, \dots$, are zero

- De manera lógica, solo se tienen en cuenta los retrasos que son múltiplos de s , por lo que los otros retrasos tendrán una autocorrelación y autocorrelación parcial nula
- Estas propiedades se pueden considerar generalizaciones de las propiedades para los modelos no estacionarios presentados anteriormente
- En general, se pueden combinar los operadores estacionales y no estacionales en un modelo multiplicativo ARMA estacional o SARMA, denotado por $ARMA(p, q) \times (P, Q)_s$, el cual tiene la siguiente forma:

$$\Phi_p(B^s)\phi(B)x_t = \Theta_Q(B^s)\theta(B)w_t$$

- Aunque las propiedades sobre la ACF y la PACF mostradas anteriormente ya no se mantengan estrictamente cuando se tiene un modelo combinado, el comportamiento de estas funciones tiene comportamientos parecidos. Para un modelo combinado como este, se tiende a ver una combinación de los hechos indicados en las tablas anteriores para modelos no estacionales y para modelos estacionales



- Cuando se ajustan estos modelos, enfocarse primero en los componentes autorregresivos y de media móvil estacionales conlleva mejores resultados
- La persistencia estacional ocurre cuando el proceso está próximo a ser periódico en la estación. En este caso, se puede pensar en modelizar la serie temporal x_t con un componente estacional S_t que sigue un camino aleatorio

$$x_t = S_t + w_t \quad \text{where} \quad S_t = S_{t-12} + v_t$$

- En este modelo, w_t y v_t son procesos de ruido blanco no correlacionados
- La tendencia de los datos a seguir este tipo de modelo se exhibirá en una ACF muestral que es grande y decae de manera muy lenta en los retrasos $h = 12k$ para $k = 1, 2, \dots$
- Si se sustraen el efecto de años sucesivos entre ellos, se puede obtener un proceso estacionario, definido a partir de la siguiente identidad:

$$(1 - B^{12})x_t = x_t - x_{t-12} = v_t + w_t - w_{t-1} = v_t + \nabla w_t$$

- Este modelo estacionario es $MA_{12}(1)$, y su ACF tendrá un pico solo en el retraso 12
- En general, la diferenciación estacional se puede indicar cuando la ACF decae lentamente en múltiplos de alguna estación s pero es ignorable entre estos periodos
 - Por lo tanto, la diferencia estacional de orden D para $D = 1, 2, \dots$ se define de la siguiente manera:

$$\nabla_s^D x_t = (1 - B^s)^D x_t$$

- Normalmente $D = 1$ es suficiente para obtener estacionariedad estacional, y permite incorporar esta idea para generalizar el modelo ARIMA
- El modelo modelo multiplicativo ARIMA estacional o SARIMA, denotado por $ARIMA(p, d, q) \times (P, D, Q)_s$, se define a partir de la siguiente ecuación:

$$\Phi_P(B^s)\phi(B)\nabla_s^D \nabla^d x_t = \delta + \Theta_Q(B^s)\theta(B)w_t$$

- En este modelo, w_t representa el ruido blanco gaussiano visto para otros modelos
- Los componentes autorregresivos y de media móvil ordinarios se representan por los polinomios $\phi(B)$ y $\theta(B)$ de orden p y q , respectivamente, mientras que los componentes estacionales se representan por $\Phi_P(B^s)$ y $\Theta_Q(B^s)$ con órdenes P y Q , respectivamente. Los componentes de diferencia ordinarios y

estacionales se representan por $\nabla^d = (1 - B)^d$ y $\nabla_s^D = (1 - B^s)^D$

- Para poder estimar y predecir este tipo de proceso, solo se tienen que hacer modificaciones bastante directas y lógicas para el caso del retraso unitario ya visto
 - En particular, la condición que se requiere es que $|\Phi| < 1$ y que $|\Theta| < 1$
 - Para hacer una predicción, no es más que utilizar los datos pasados y usarlos para obtener el valor de m pasos adelante deseados

SARIMA(0,1,1)(0,1,1)₁₂ :

$$\tilde{x}_{n+1} = x_n + (x_{n-11} - x_{n-12}) + \tilde{w}_{n+1} + \theta_1 w_n + \theta_{12} w_{n-11}$$

$$+ \theta_1 \theta_{12} w_{n-12}$$

$$\tilde{x}_{n+2} = \tilde{x}_{n+1} + (x_{n-10} - x_{n-11}) + \tilde{w}_{n+2} + \theta_1 \tilde{w}_{n+1} + \theta_{12} w_{n-10}$$

$$+ \theta_1 \theta_{12} w_{n-11}$$

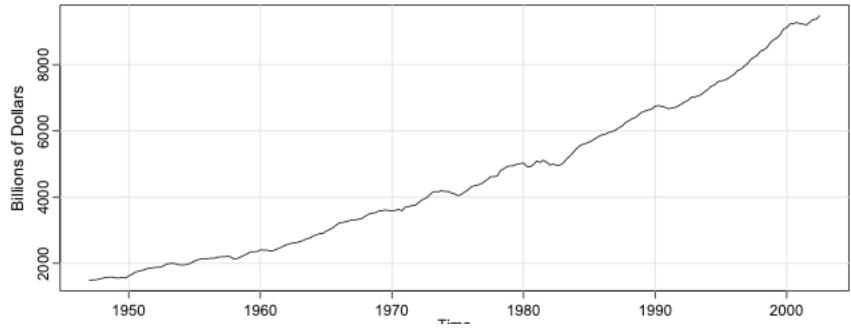
...

$$\Phi_p(B^s) \phi(B) \nabla_s^D \nabla^d x_{n+m} = \delta + \Theta_Q(B^s) \theta(B) w_{n+m}$$

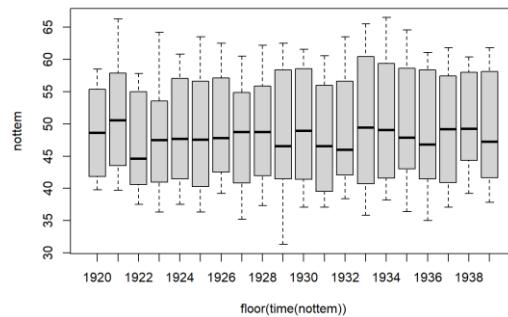
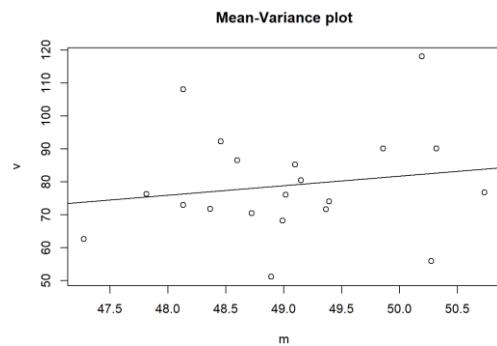
- Debido a que seleccionar el modelo SARIMA apropiado para un conjunto de datos es difícil, primero se tiene que identificar el modelo, analizar los residuos y después pasar a la fase de diagnóstico
 - Se puede pensar primero en términos de encontrar los operadores de diferencia que produzcan una serie más o menos estacionaria para poder identificar después un modelo ARMA o SARMA para ajustar las series de residuos

Los modelos ARIMA: la identificación y el diagnóstico

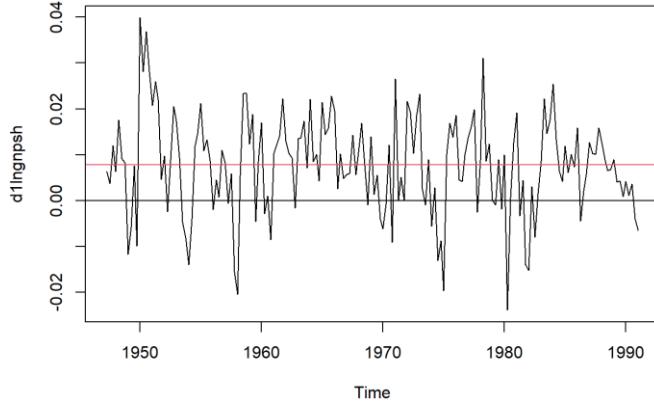
- Existen algunos pasos básicos para identificar modelos ARIMA (y ARMA) para datos de series temporales, los cuales involucran graficar los datos, transformarlos, identificar los órdenes de dependencia del modelo
 - Primero, igual que con cualquier análisis de datos, se tiene que construir un gráfico para los datos e inspeccionar el gráfico para detectar anomalías



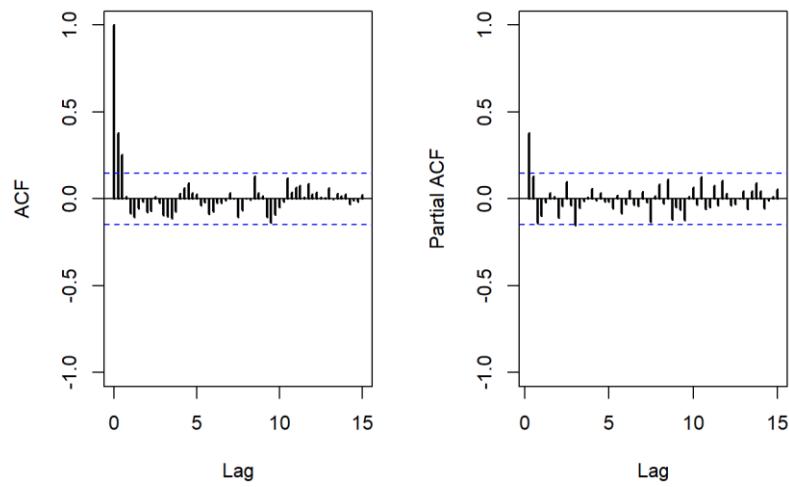
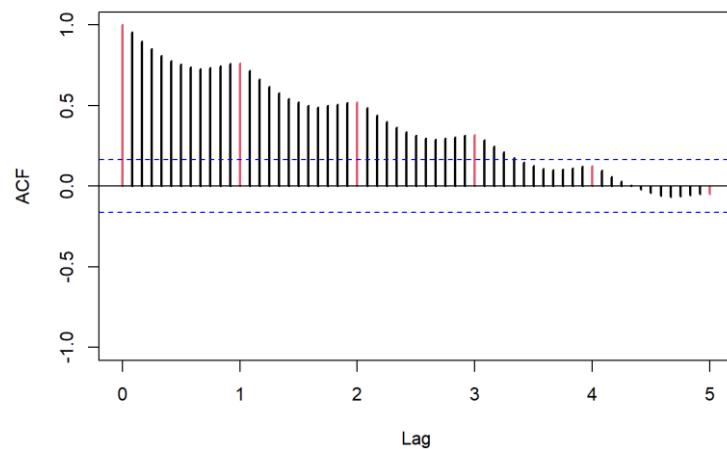
- Con el gráfico es posible identificar el comportamiento de la variabilidad en los datos, aunque otros gráficos como el de la media contra la varianza o el de cajas. A partir de estos gráficos, se decide si aplicar o no una transformación Box-Cox, recordando que lo importante es ver la constancia de la varianza, no los diferentes niveles (para los gráficos de caja)



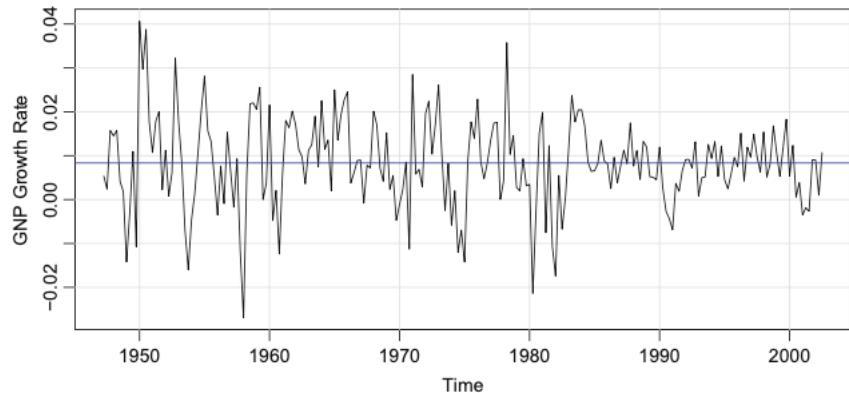
- También es posible detectar tendencias en los datos a través del gráfico de la serie original o transformada, por lo que se puede decidir si diferenciar o eliminar la tendencia directamente (a través de un modelo polinómico). Si la media es diferente a cero, esta se tiene que restar a toda la serie para poder trabajar con una serie con media nula



- Después de transformar los datos, el siguiente paso es identificar los valores preliminares de orden autorregresivo p , el orden de diferenciación d , y el orden de media móvil q
 - Lo primero de todo es, por supuesto, comprobar que la serie transformada no es ya una serie estacionaria de por si. Esto se puede comprobar a través de los gráficos de la ACF y la PACF, en donde la gran mayoría de los retrasos no tendrían que ser significativos o no estar muy lejos de las bandas de confianza



- Como se ha mencionado, el gráfico permite identificar tendencias, de modo que se puede inspeccionar el gráfico para las series diferenciadas y así escoger el orden d



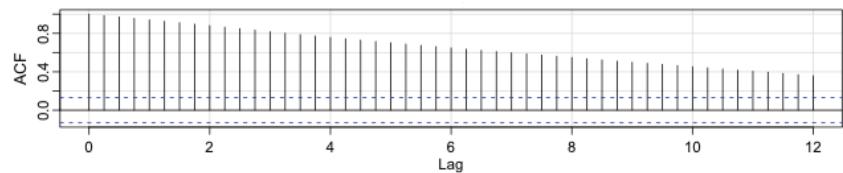
- El orden d se restará al orden del polinomio $\phi(B)$ (a p), por lo que el modelo resultante sería un $ARIMA(p - d, d, q)$, el cual se puede expresar de la siguiente manera:

$$\phi(B)\nabla^d x_t = \theta(B)w_t$$

$$\phi(B) = (1 - \phi_1 B - \cdots - \phi_p B^p) =$$

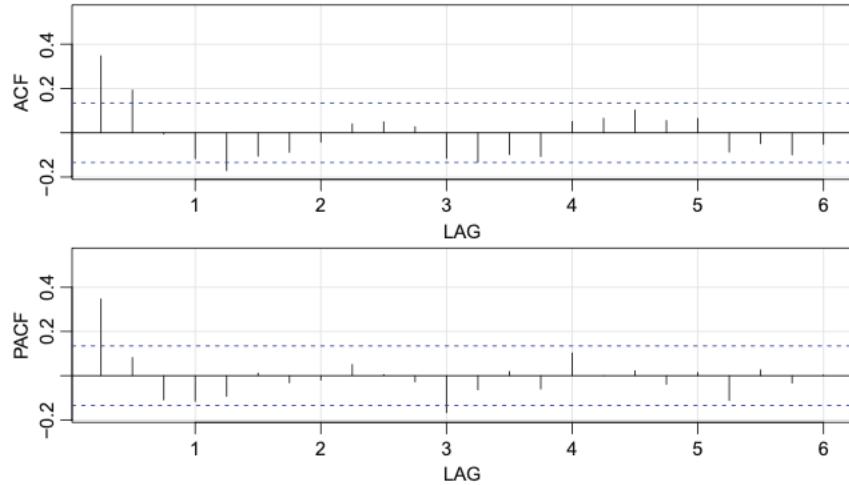
$$= (1 - \phi'_1 B - \cdots - \phi'_{p-d} B^{p-d})(1 - B)^d$$

- No obstante, siempre se tiene que tener cuidado para no sobrediferenciar e introducir dependencia inexistente (ver cuando la varianza incremente). Por lo tanto, una manera de poder ver el orden de diferenciación d es diferenciando hasta el punto en el que un mayor d cause el aumento de la varianza del modelo
- El gráfico de la ACF muestral también permite identificar cuando diferenciar es necesario: como el polinomio $\phi(z)(1 - z)^d$ tiene una raíz unitaria, $\hat{\rho}(h)$ no decaerá hacia cero rápido mientras h incrementa, por lo que una caída lenta indicaría la necesidad de diferenciación

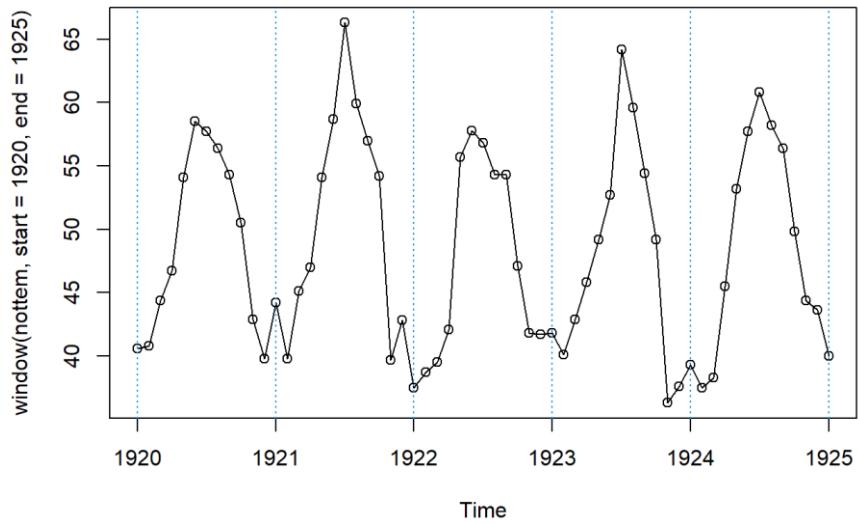


- Una vez se ha determinado d , se puede mirar la ACF y la PACF muestrales de $\nabla^d x_t$ y así obtener los valores para p y q

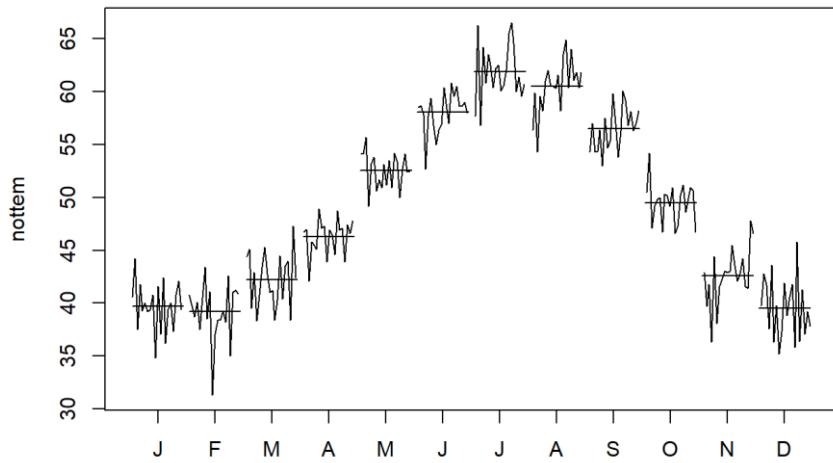
dependiendo de los criterios vistos en la tabla anteriormente. Se tienen que considerar valores de p y q lo más pequeños posibles, dado el principio de parsimonia y el sentido que esto tendría dependiendo de la naturaleza de la serie temporal



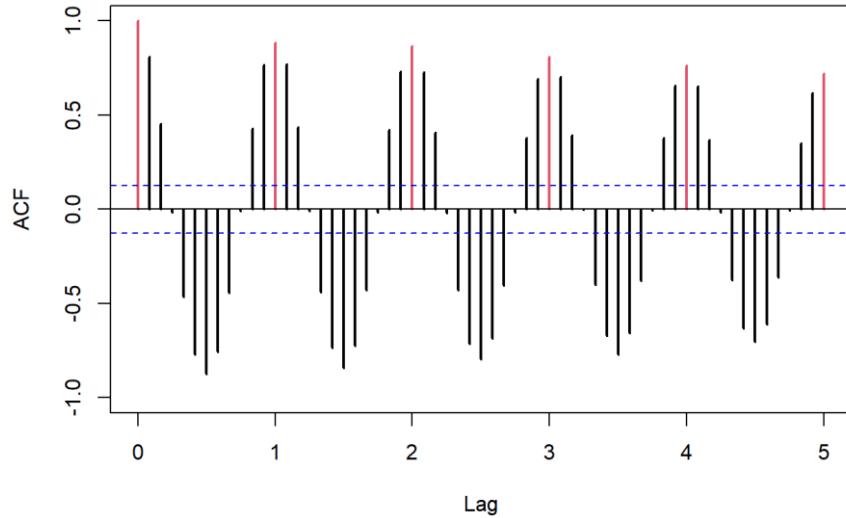
- Debido a que hay veces que es difícil ver qué gráfico es el que corta y cual es el que tiende a cero (los dos no pueden nunca cortar a la vez), además de situaciones en donde dos modelos pueden ser muy similares, aunque sean diferentes, se suelen proponer varios modelos o alternativas para p y q . Con estas elecciones para p , d y q , es posible comenzar a estimar parámetros
- Si no se está muy seguro del modelo, pero tanto la ACF como la PACF decaen, entonces se puede escoger un modelo ARMA(1,1) como modelo preliminar
- Para identificar modelos SARIMA se suelen seguir pasos similares a los vistos anteriormente para modelos ARIMA, pero con algunas diferencias debidas a la estructura estacional
 - Primero se tiene que construir un gráfico para los datos e inspeccionar el gráfico para detectar anomalías
 - Antes que nada, se tienen que llevar a cabo los diferentes pasos mostrados anteriormente para los modelos ARIMA, de modo que la revisión de la parte regular esté ya terminada
 - Una vez hechos, se tienen que identificar patrones estacionales en la serie temporal. Para ello, el primer gráfico que se puede ver es el de la misma serie, de modo que, si se puede identificar fácilmente un patrón repetitivo, se puede concluir que habría una estructura estacional



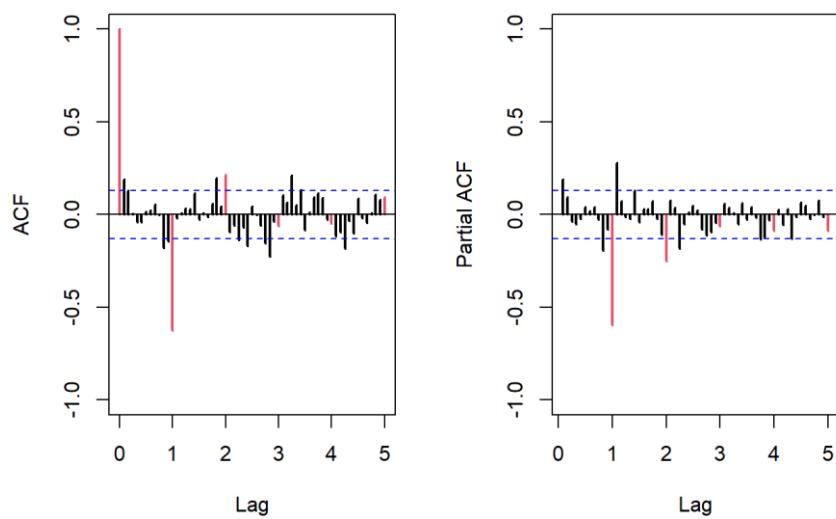
- De manera mucho más clara, se puede usar un *monthplot* (un gráfico en el que se considera la serie temporal solo para observaciones en una frecuencia concreta) para poder ver si la media de cada día, mes u medida de tiempo es la misma (no hay patrón estacional) o es diferente (indicando estacionalidad)



- Si se identifica la presencia de una estructura estacional, entonces la serie no será estacionaria (la densidad de probabilidad cambia) y por tanto será necesario utilizar la diferenciación estacional ($1 - B^s$), en donde s es la frecuencia de los datos
- Para poder saber si la serie se vuelve estacionaria al aplicar la diferenciación, se tiene que comprobar la significación de los retrasos en los retrasos correspondientes a la frecuencia de la serie temporal (en el ejemplo, aquellos retrasos rojos). Si se ve un patrón de decrecimiento muy lento, entonces la serie no es estacionaria

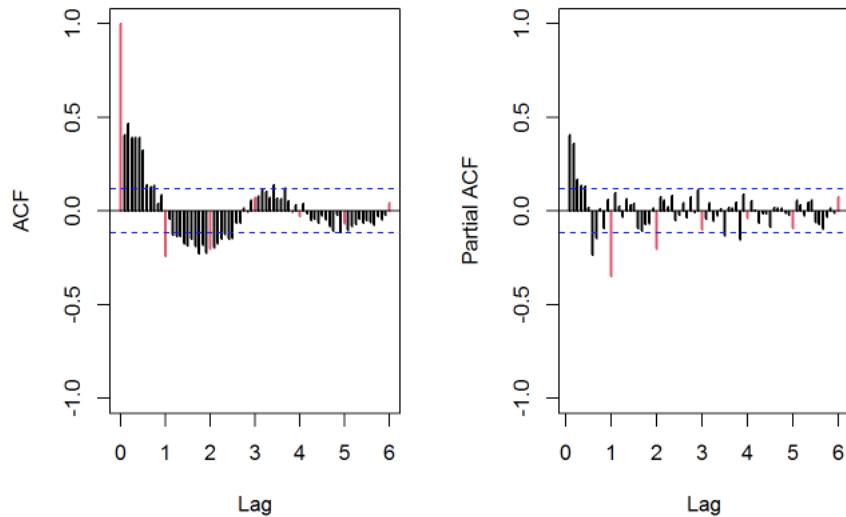


- Si, en cambio, se puede ver un decrecimiento de los retrasos que corresponden a la frecuencia de la serie, entonces la serie será estacionaria



- Los otros retrasos (los que no corresponden a la frecuencia de los datos) tendrían que verse más o menos estacionarios debido a que se asume que ya se ha hecho la diferenciación para la parte regular
 - Una vez la serie se ha hecho estacionario tanto por la parte regular como para la parte estacional, es posible identificar los órdenes regulares y estacionales para el SARIMA
 - En esta parte de la identificación, los patrones anteriormente descritos se aplican, de modo que se tiene que ver el comportamiento de la ACF y de la PACF teniendo en cuenta los retrasos que corresponden a la frecuencia de los datos y aquellos que no

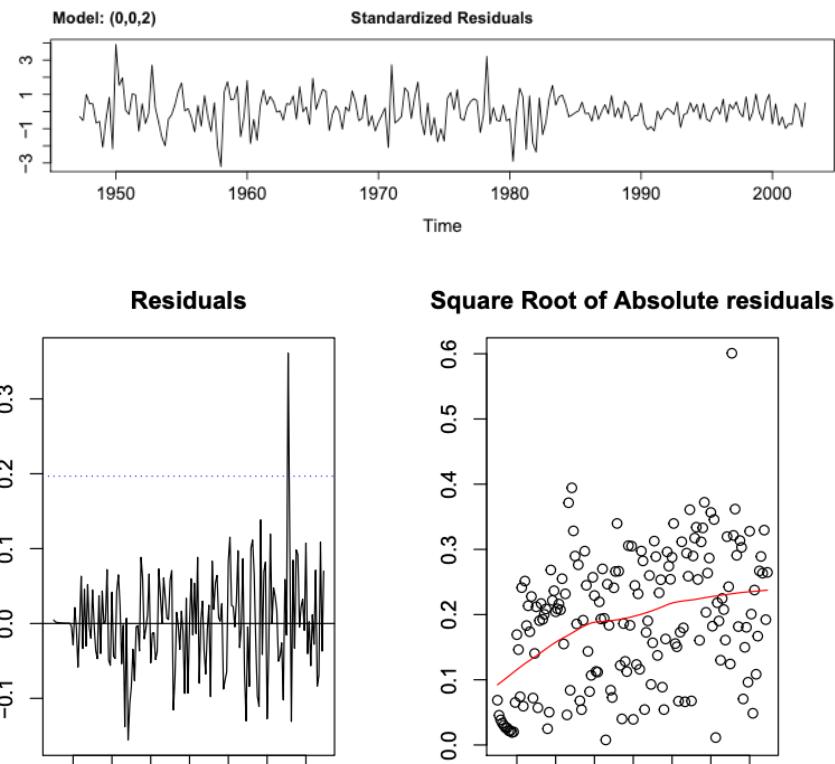
- Normalmente, uno se fija primero en la parte de los retrasos estacionales, de modo que primero se tiene que encontrar el orden de la parte estacional



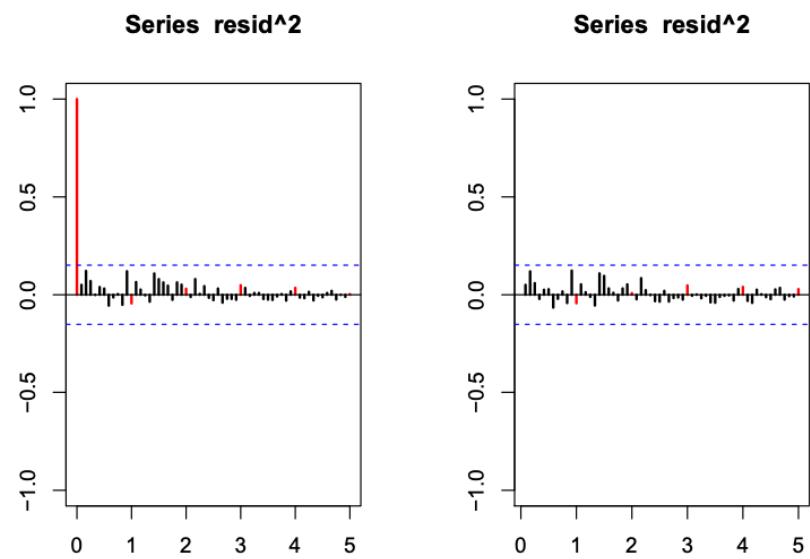
- Despues de escoger el orden estacional, se mira los retrasos que no son estacionales, de modo que se escoge como con un modelo ARMA
- Si cuando se ajustan los modelos se ve que un coeficiente no es significativo para la parte estacional, entonces se necesita reducir el orden de este, igual que con los coeficientes para los modelos ARMA
- Una vez se ha identificado el modelo y se ha ajustado, se puede hacer un diagnóstico de este a partir de los residuos. Para ello, se realizan varios pasos como los siguientes:
 - Lo primero es hacer un gráfico de las innovaciones o residuos $x_t - \hat{x}_t^{t-1}$ o las innovaciones estandarizadas, en donde \hat{x}_t^{t-1} es la predicción a un paso adelante de x_t basada en el modelo ajustado y \hat{P}_t^{t-1} es la varianza del error estimada de un paso adelante

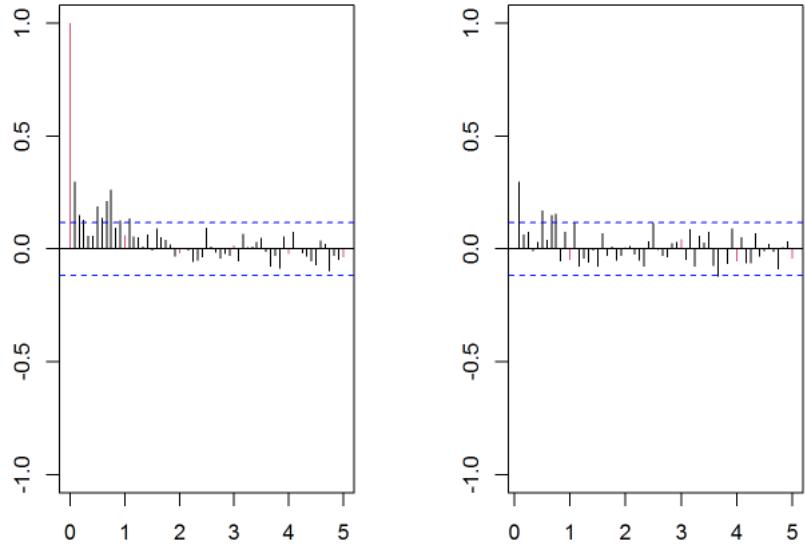
$$e_t = \frac{x_t - \hat{x}_t^{t-1}}{\sqrt{\hat{P}_t^{t-1}}}$$

- Si el modelo se ajusta bien, los residuos estandarizados deberían comportarse como una secuencia iid con media cero y varianza unitaria. El gráfico de los residuos estandarizados y el de la raíz cuadrada de los residuos absolutos permiten inspeccionar cualquier desviación obvia de este comportamiento (por ejemplo, en donde la varianza no es constante y crece/decrece)

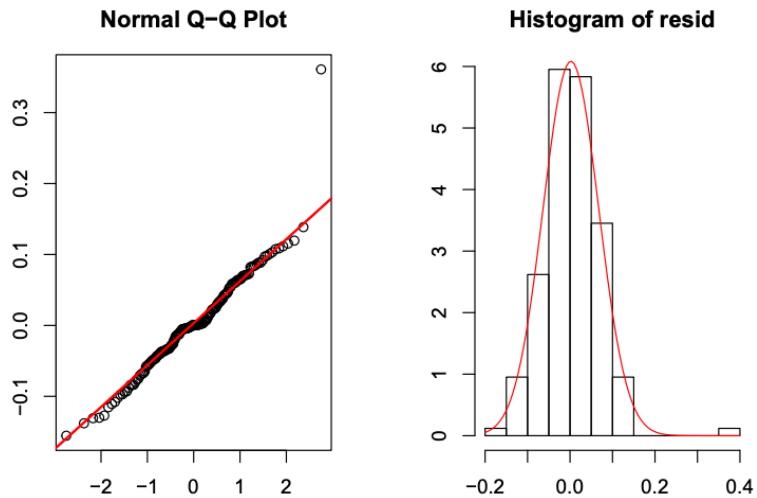


- También se pueden utilizar la ACF y la PACF de los residuos cuadrados para poder identificar la predictibilidad de la volatilidad, de modo que, si no hay un comportamiento propio de un ruido blanco, entonces la suposición de homocedasticidad no se mantiene

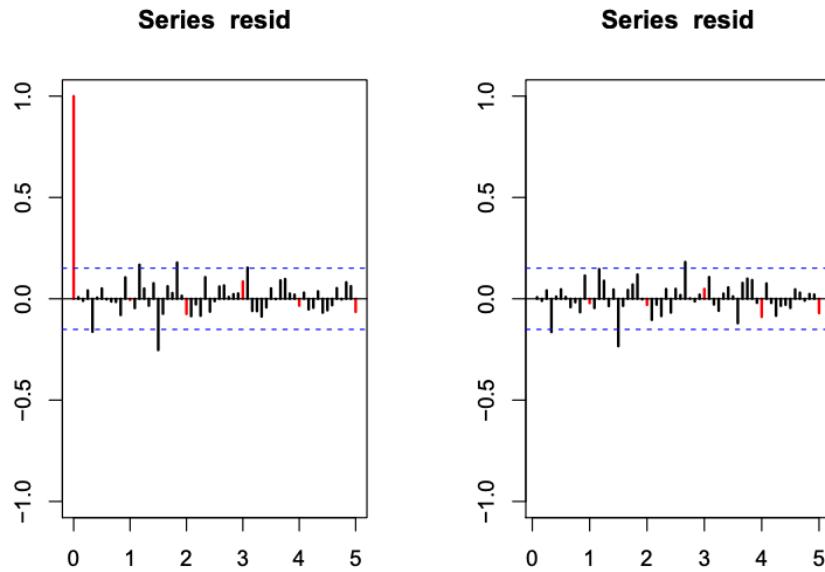




- Un contraste estadístico que permite comprobar la suposición de homocedasticidad del modelo es el de Breusch-Pagan, el cual tiene como hipótesis nula la homoscedasticidad del modelo. En el caso de rechazar la hipótesis nula, se estaría favoreciendo a la heteroscedasticidad
- En este paso es posible detectar valores atípicos (a través del histograma) y volatilidad o varianza no constante (a través del gráfico de residuos), por lo que se tiene que realizar un tratamiento de los valores atípicos y/o usar modelos para la varianza (como los GARCH)
- Si la serie temporal no es normal, no es suficiente con que los residuos sean independientes, por lo que se tiene que investigar la normalidad de los residuos (además, la mayoría de veces se supone que el ruido blanco es normal)
 - La investigación de la normalidad de los errores se puede hacer a través de un histograma de los residuos y de un *Q-Q plot*



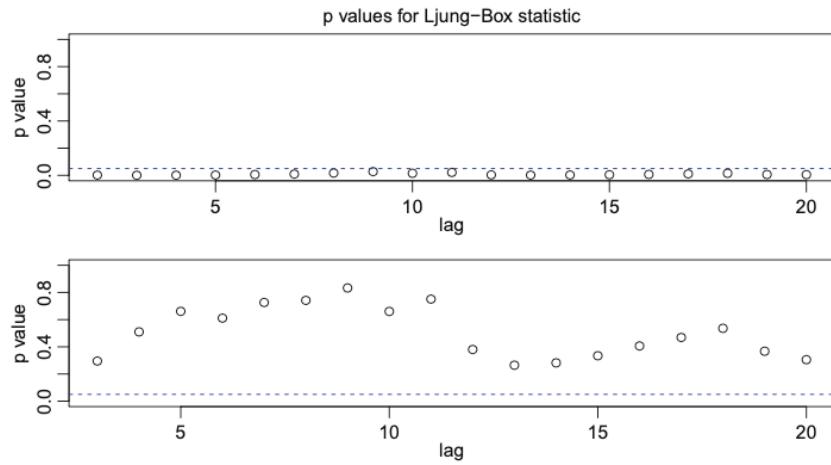
- También se pueden utilizar contrastes de normalidad univariante como el contraste de Shapiro-Wilks o de normalidad multivariante, tales como el de Jarque-Bera o el de Anderson-Darling. En todos estos contrastes, la hipótesis nula es que los datos se distribuyen de manera normal, de modo que rechazar esta implica no normalidad
- En este paso es posible detectar valores atípicos (los cuales se tendrán que tratar), asimetría o bimodalidad (lo cual requerirá una transformación, la eliminación de valores atípicos o el cambio de distribución de los residuos) y exceso de curtosis (por lo que se tienen que utilizar modelos de volatilidad o distribuciones como la t-Student)
- Para poder investigar la aleatoriedad o independencia de los residuos, es posible utilizar contrastes de aleatoriedad o, más fácilmente, observar las autocorrelaciones muestrales de los residuos para detectar cualquier patrón o valores anormalmente grandes
- Para una secuencia de ruido blanco, las autocorrelaciones muestrales son aproximadamente independientes y normalmente distribuidas con media nula y varianzas $1/n$. Por lo tanto, para evaluar la estructura de autocorrelación de los residuos, se mira el gráfico de la ACF junto a bandas de confianza de $\pm 1.96/\sqrt{n}$



- Los residuos de un ajuste de modelo no se comportarán perfectamente como un ruido blanco, y la varianza de $\hat{\rho}_e(h)$ será mucho menor que $1/n$. Se considera que no hay dependencia en el modelo si se ve que solo los valores de $\hat{\rho}_e(h)$ para h muy alejadas son significativos (la lejanía depende de la frecuencia de los datos)
- Además, también es posible utilizar un contraste general que tenga en cuenta las magnitudes de $\hat{\rho}_e(h)$ en un grupo, como el contraste usando el estadístico Q de Ljung-Box-Pierce (comúnmente con $H = 20$), en donde la hipótesis nula es que el modelo es adecuado

$$Q = n(n+2) \sum_{h=1}^H \frac{\hat{\rho}_e^2(h)}{n-h} \sim \chi_{H-p-q}^2$$

- Si w_t es un ruido blanco, entonces $n\hat{\rho}_w^2(h)$ para $h = 1, 2, \dots, H$ tienen que ser variables aleatorias χ_1^2 asintóticamente independientes, por lo que $n \sum_{h=1}^H \hat{\rho}_e^2(h) \sim \chi_H^2$. Como el contraste requiere el ACF de los residuos de un ajuste del modelo, hay una pérdida de $p + q$ grados de libertad (para el AR y el MA), por lo que se restan estos grados de libertad
- Este contraste producirá unos *p-values* que permitirán evaluar, para cada h , si hay independencia. Si hay algún valor por debajo de la línea marcada (al nivel α), entonces se rechaza la hipótesis nula de adecuación del modelo o de independencia



- Este contraste producirá unos *p-values* que permitirán evaluar, para cada h , si hay independencia. Si hay algún valor por debajo de la línea marcada (al nivel α), entonces se rechaza la hipótesis nula de adecuación del modelo o de independencia para ese retraso h , pero no se rechaza la hipótesis de manera conjunta para los otros retrasos
- Si se detecta que hay retrasos significativos, entonces se tiene que reidentificar el modelo o añadir parámetros en este
- Una vez se han validado las suposiciones de los modelos propuestos a través del estudio de los residuos, se puede seleccionar el modelo a través de diferentes criterios
 - Los criterios más utilizados son los criterios de información, tales como el de Bayes y el de Akaike, ya que balancean la bondad del ajuste con la simplicidad del modelo
 - En este caso, se escogerá el modelo con un menor valor del criterio de información
 - Otros criterios como la verosimilitud logarítmica o la s^2 estimada no sirven para comparar modelos, dado que no tienen en cuenta la complejidad de estos
 - También es posible investigar la causalidad y la invertibilidad del modelo ajustado, a través de calcular las raíces de los polinomios característicos del modelo
 - Esto se puede hacer a través de igualar los polinomios a cero (cada uno de ellos por separado) y obtener las raíces. Una vez hecho esto, se comprueba el valor absoluto o el módulo (si es un número complejo)

- Finalmente, se pueden utilizar algunas medidas con tal de hacer una evaluación de la capacidad predictiva del modelo ajustado
 - El porcentaje de error absoluto medio o *mean average percentage error* (MAPE) es una medida de la exactitud o *accuracy* de las predicciones (que tan cerca están del valor verdadero)

$$MAPE = \frac{100}{n} \sum_{t=1}^n \left| \frac{x_t - \hat{x}_t}{x_t} \right|$$

- La raíz cuadrada del porcentaje del error cuadrático medio o *root mean squared percentage error* (RMSPE) es una medida de la precisión de las predicciones (que tan cerca están las predicciones de un valor entre ellas)

$$RMSPE = \sqrt{\frac{1}{n} \sum_{t=1}^n \left(\frac{x_t - \hat{x}_t}{x_t} \right)^2}$$

- La longitud del intervalo de confianza o de las bandas de confianza para las predicciones también permite obtener una idea de la dispersión de las predicciones

Los modelos ARIMAX: análisis de intervención y de valores atípicos

- Hay veces que el fenómeno que se mide en una serie temporal x_t se ve afectada por factores exógenos que pueden cambiar las dinámicas naturales del proceso
 - Existen varias series temporales reales en las que pueden ocurrir eventos exógenos que afecten de manera significativa la dinámica del proceso

Time series	Fact
Daily number of cars in a highway	A serious accident
Monthly number of passengers in an airline	A terrorist attack on a plane
Monthly amount of production in a factory	A two-weeks strike
Weekly consumption of gasoline	An important increasing in the price
Monthly passengers in Subway	An epidemic outbreak
Number of daily reported infected people	A change in the methodology

- Algunos de estos hechos se pueden conocer con antelación, pero para el tratamiento de valores atípicos, se considera que los eventos son inesperados

- Estos hechos pueden afectar a la serie temporal de manera que el próximo valor observado no es el valor esperado (condicional) correspondiente al comportamiento natural de la serie
 - La afectación puede extenderse a varios periodos en el tiempo con diferente intensidad
- Desde un punto de vista estadístico, la presencia de valores atípicos tiene efectos como los siguientes:
 - Puede cambiar la estructura de la autocorrelación (a través de la ACF y la PACF)
 - Puede cambiar el valor y/o la significación de los parámetros del modelo
 - Puede incrementar la varianza residual y puede invalidar el modelo (a partir del efecto en la homocedasticidad, la normalidad y la independencia de los residuos)
- Como uno está interesado en la predicción, estos efectos harían que el modelo diera predicciones inexactas e imprecisas si se incluyen valores atípicos, por lo que hay que tratarlos
 - La idea detrás del tratamiento de valores atípicos es identificar el momento en el que el evento afectó a las series, medir el impacto en las observaciones alteradas, reconstruir las series y aplicar la metodología de Box-Jenkins descrita anteriormente
 - Para poder incluir el efecto del valor atípico en las series, es necesario incluir una variable exógena a la serie linealizada teórica, obteniendo así la serie observada
 - La estimación de los parámetros para esta parte exógena puede ser útil para obtener la serie teórica, que sería la serie temporal que se observaría si el evento no hubiera ocurrido)
- Para poder tener en cuenta efectos externos al analizar series temporales, se utiliza el modelo ARMAX, el cual es un modelo de regresión lineal que tiene en cuenta variables exógenas a parte de la serie observada
 - En mínimos cuadrados ordinarios, se asume que x_t es un ruido gaussiano, de modo que $\gamma_x(s, t) = 0$ para $s \neq t$ y $\gamma_x(t, t) = \sigma^2$, independiente de t
 - Si este no es el caso, entonces se necesita utilizar los mínimos cuadrados ponderados

- Escribiendo el modelo de forma vectorial, en donde $\boldsymbol{\Gamma} = \{\gamma_x(s, t)\}$ es una matriz $n \times n$ simétrica, se puede multiplicar cada componente por $\boldsymbol{\Gamma}^{1/2}$

$$\boldsymbol{\Gamma}^{1/2}\mathbf{y} = \boldsymbol{\Gamma}^{1/2}\mathbf{Z}\boldsymbol{\beta} + \boldsymbol{\Gamma}^{1/2}\mathbf{x} \Rightarrow \mathbf{y}^* = \mathbf{Z}^*\boldsymbol{\beta} + \boldsymbol{\delta}$$

- En este caso $\mathbf{y} = (y_1, \dots, y_n)'$ y $\mathbf{x} = (x_1, \dots, x_n)'$ son vectores $n \times 1$, $\boldsymbol{\beta} = (\beta_1, \beta_2, \dots, \beta_r)'$ es de tamaño $r \times 1$ y $\mathbf{Z} = [\mathbf{z}_1 | \mathbf{z}_2 | \dots | \mathbf{z}_n]'$ es la matriz $n \times r$ compuesta de las variables de insumo. Consecuentemente, la matriz de varianzas y covarianzas de $\boldsymbol{\delta}$ es la matriz identidad (se escala) y el modelo es el modelo lineal clásico
- El estimador ponderado de $\boldsymbol{\beta}$ y de la matriz de varianzas y covarianzas $Cov(\widehat{\boldsymbol{\beta}})$ serían los siguientes:

$$\widehat{\boldsymbol{\beta}}_w = (\mathbf{Z}^{*'}\mathbf{Z}^*)^{-1}\mathbf{Z}^{*'}\mathbf{y}^* = (\mathbf{Z}'\boldsymbol{\Gamma}^{-1}\mathbf{Z})^{-1}\mathbf{Z}'\boldsymbol{\Gamma}^{-1}\mathbf{y}$$

$$Cov(\widehat{\boldsymbol{\beta}}_w) = (\mathbf{Z}'\boldsymbol{\Gamma}^{-1}\mathbf{Z})^{-1}$$

- Si y_t es ruido blanco, entonces $\boldsymbol{\Gamma} = \sigma^2\mathbf{I}$ y estos resultados se reducen a los de mínimos cuadrados ordinarios
- En el caso de las series temporales, normalmente es posible asumir una estructura de covarianza estacionaria para el proceso de error x_t que corresponde a un proceso lineal e intentar encontrar una representación ARMA para y_t
 - Si se tiene un error $AR(p)$ puro, entonces $\phi(B)x_t = w_t$ es la transformación lineal que, cuando se aplica el proceso de error, se produce un proceso de ruido blanco w_t . Multiplicando la ecuación de regresión a través de la transformación $\phi(B)$ da una regresión lineal de la siguiente forma:

$$\phi(B)y_t = \sum_{j=1}^r \beta_j \phi(B)z_{tj} + \phi(B)x_t \Rightarrow y_t^* = \sum_{j=1}^r \beta_j z_{tj}^* + w_t$$

- En el caso AR se tiene que resolver un problema de mínimos cuadrados para minimizar el error de la suma de errores al cuadrado

$$S(\phi, \beta) = \sum_{t=1}^n w_t^2 = \sum_{t=1}^n \left[\phi(B)y_t - \sum_{j=1}^r \beta_j \phi(B)z_{tj} \right]^2$$

- Si se tiene un error $ARMA(p, q)$, entonces $\phi(B)x_t = \theta(B)w_t$ y se transforma a $\pi(B)x_t = w_t$, en donde $\pi(B) = \theta(B)^{-1}\phi(B)$. En este caso, la suma de errores cuadrados será la siguiente:

$$\pi(B)y_t = \sum_{j=1}^r \beta_j \pi(B)z_{tj} + \pi(B)x_t \Rightarrow y_t^* = \sum_{j=1}^r \beta_j z_{tj}^* + x_t$$

$$\Rightarrow S(\phi, \beta) = \sum_{t=1}^n w_t^2 = \sum_{t=1}^n \left[\pi(B)y_t - \sum_{j=1}^r \beta_j \pi(B)z_{tj} \right]^2$$

- En este punto, el problema principal es que normalmente no se sabe el comportamiento del ruido y_t antes del análisis, de modo que se puede llevar a cabo el siguiente algoritmo para solucionar el problema:
 - Primero se estima una regresión ordinaria de y_t sobre $z_{t1}, z_{t2}, \dots, z_{tr}$ (actuando como si los errores no estuvieran correlacionados). En este punto se retienen los residuos, $\hat{x}_t = y_t - \sum_{j=1}^r \hat{\beta}_j z_{tj}$
 - Despues se identifica el modelo ARMA para los residuos \hat{x}_t
 - Estimar los mínimos cuadrados ponderados (o MLE) en el modelo de regresión con errores autocorrelacionados usando el modelo especificado en el segundo paso
 - Inspeccionar los residuos \hat{w}_t para ver si son un proceso de ruido blanco y ajustar el modelo si es necesario
- Siendo y_t la serie observada original, z_{tj} las variables exógenas para $j = 1, 2, \dots, r$ e x_t la serie teórica sin los efectos externos, el modelo ARMAX se deriva a partir de la regresión con errores autocorrelacionados

$$y_t = \sum_{j=1}^r \beta_j z_{tj} + x_t \Rightarrow x_t = y_t - \sum_{j=1}^r \beta_j z_{tj}$$

- A partir de este modelo para las series sin efectos, entonces se pueden aplicar los pasos anteriores para poder obtener modelos estacionarios

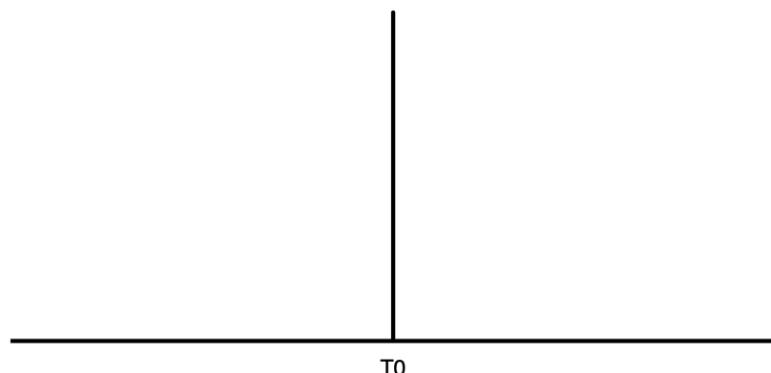
$$\phi(B)y_t = \theta(B)w_t$$

- El modelo ARIMAX se puede utilizar para hacer análisis de intervención, el cual se basa en estimar el efecto que tiene un factor externo sobre la serie temporal de interés

- En este caso, se tiene que modelar la intervención como una variable exógena, a través de una función de transferencia adecuada y una serie auxiliar de esa función de transferencia (para poder usarla como una variable exógena en el modelo)

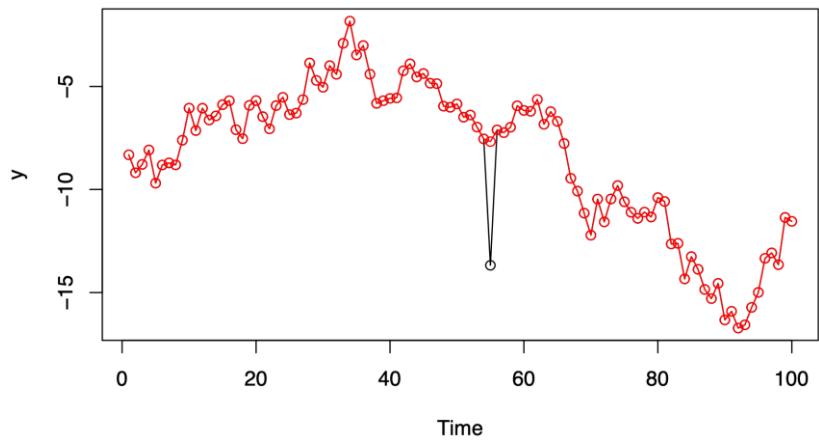
	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
2006	0	0	0	0	0	0	0	0	0	0	0	0
2007	0	0	0	0	0	0	0	0	0	0	0	0
2008	0	0	0	0	0	0	0	0	0	0	0	0
2009	0	0	0	0	0	0	1	1	1	1	1	1
2010	1	1	1	1	1	1	1	1	1	1	1	1
2011	1	1	1	1	1	1	1	1	1	1	1	1

- Una vez se modela la intervención a través de una función de transferencia adecuada y con una serie auxiliar, se puede hacer la estimación del modelo ARIMAX y hacer un contraste de hipótesis sobre la significación del parámetro correspondiente, con tal de comprobar si la intervención no ha sido efectiva (H_0) o sí
- Los tres tipos de funciones de transferencia para valores atípicos o efectos externos que se consideran normalmente son el valor atípico aditivo, el cambio transitorio y el desplazamiento de nivel
 - El valor atípico aditivo es aquel que solo tiene efecto en un periodo concreto. Por lo tanto, esto se puede modelar a través de una función de transferencia de pulso

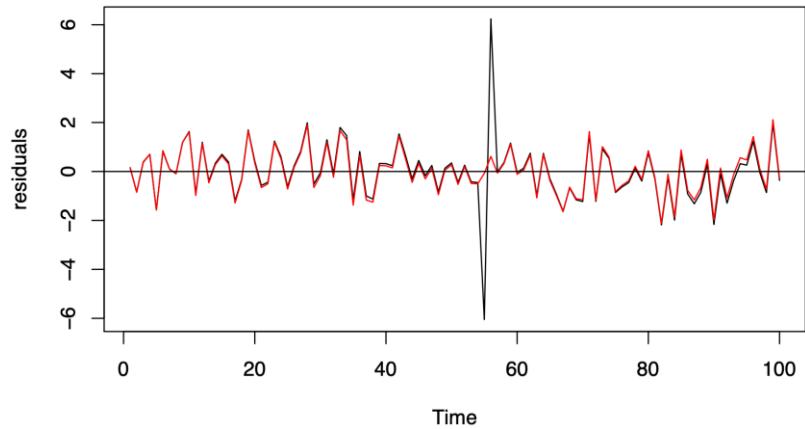


$$z_t = \mathbf{1}_{t=t_0}(t)$$

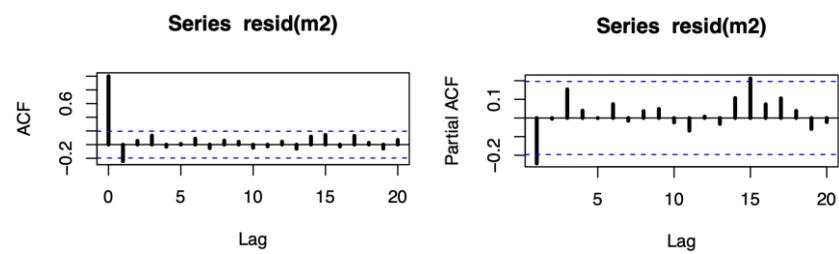
- Cuando existen este tipo de valores atípicos, se puede ver como la serie observada (color negro) y la serie lineal teórica (color rojo) son iguales excepto por una observación en t_0

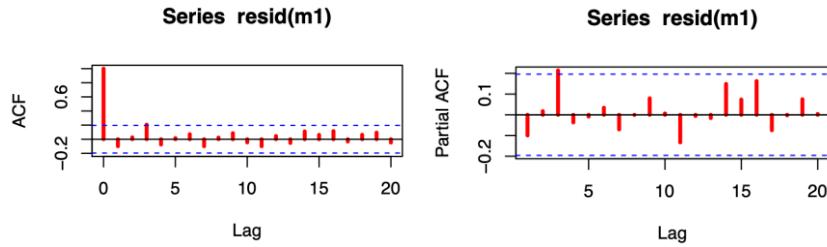


- Se puede ver como los residuos de la serie teórica linealizada estarán más en sintonía con los residuos de la serie en otros puntos en el tiempo, mientras que hay un salto en la de la serie observada (sin incluir la variable exógena)

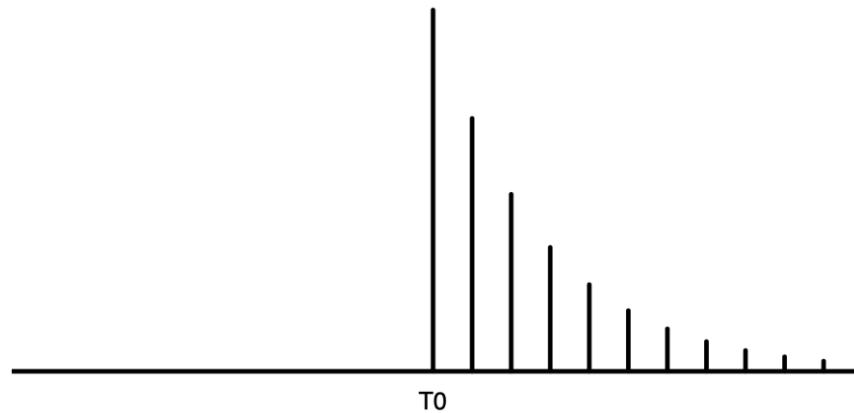


- Si se comparan la ACF y la PACF de la serie observada con la del modelo lineal teórico, se puede ver como hay diferencias entre ambos modelos (por la inclusión de la función de transferencia en el modelo). Ahora, el modelo teórico permitirá detectar mejor cuál sería el modelo para la serie temporal sin perturbar



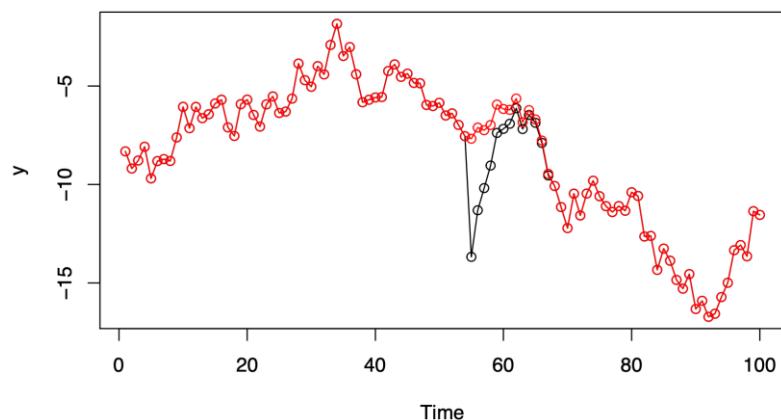


- El cambio transitorio es aquel que tiene efecto en un periodo, pero este decrece en los siguientes periodos. Por lo tanto, esto se puede modelar a través de una función de transferencia exponencial decreciente con tasa δ



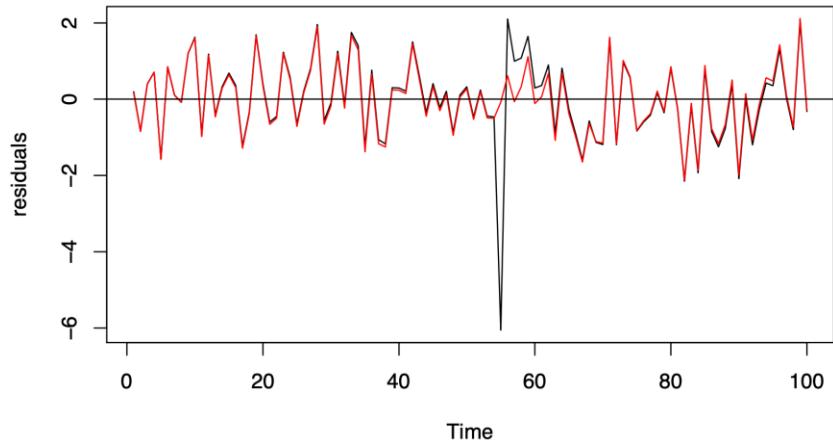
$$z_t = \delta^{(t-t_0)} \mathbf{1}_{t \geq t_0}(t) \approx \frac{1}{1-\delta} \mathbf{1}_{t=t_0}(t)$$

- Cuando existen estos tipos de valores atípicos, se puede ver como la serie observada y la serie lineal teórica son iguales excepto por las observaciones t_0 y aquellas vecinas después de t

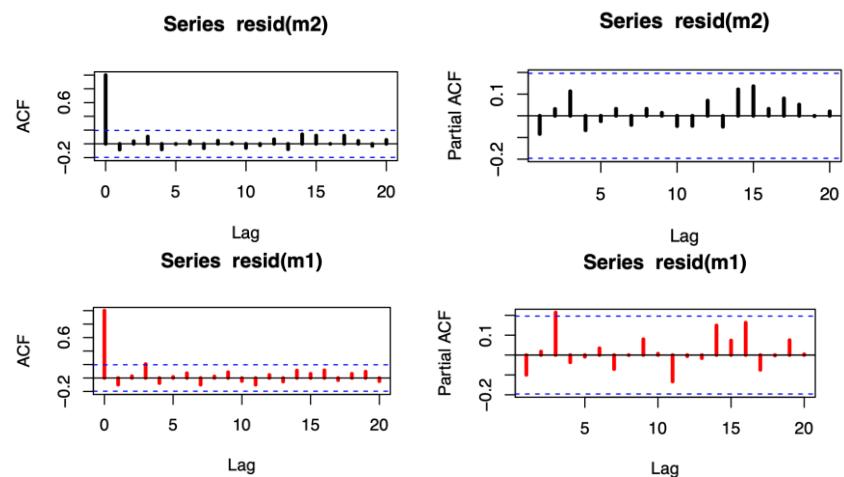


- Se puede ver como los residuos de la serie teórica linealizada estarán más en sintonía con los residuos de la serie en otros puntos en el tiempo, mientras que hay un salto y diferencias que van decreciendo para la serie original (sin incluir la variable

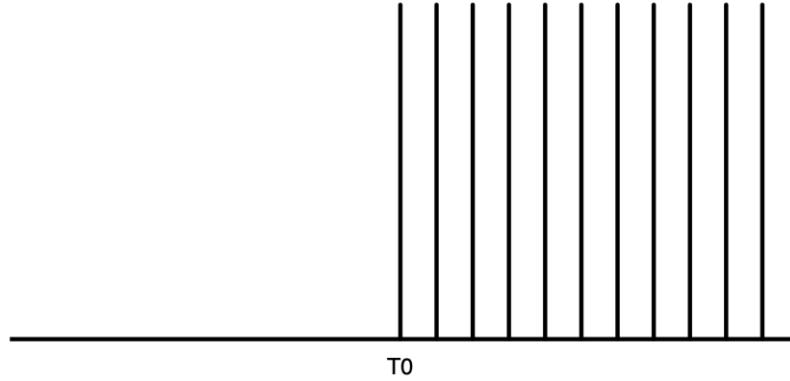
exógena), ya que difiere durante momentos posteriores a la línea del otro modelo, pero recobra la sintonía en un punto



- Si se comparan la ACF y la PACF de la serie observada con la del modelo lineal teórico, se puede ver como hay diferencias entre ambos modelos (por la inclusión de la función de transferencia en el modelo). Ahora, el modelo teórico permitirá detectar mejor cuál sería el modelo para la serie temporal sin perturbar

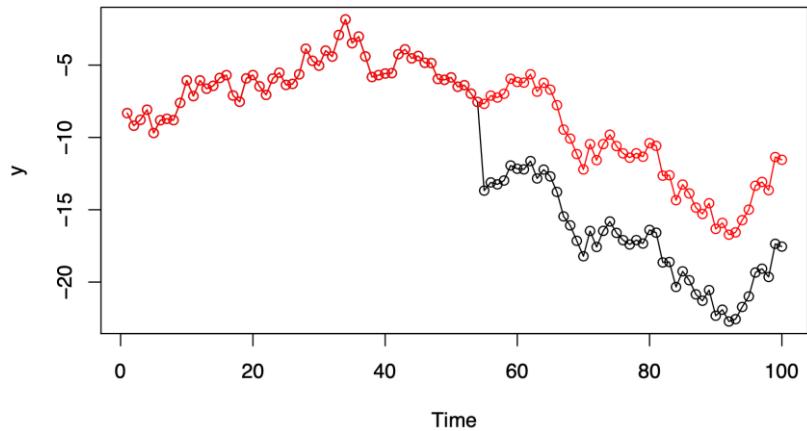


- El desplazamiento de nivel es aquel que tiene efecto en un periodo, pero que se mantiene para los siguientes periodos. Por lo tanto, esto se puede modelar a través de una función de transferencia de pasos

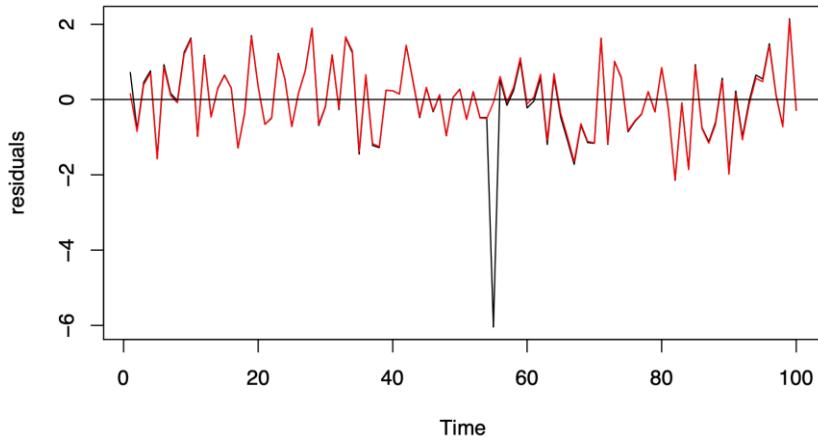


$$z_t = \mathbf{1}_{t \geq t_0}(t) \approx \frac{1}{1 - B} \mathbf{1}_{t=t_0}(t)$$

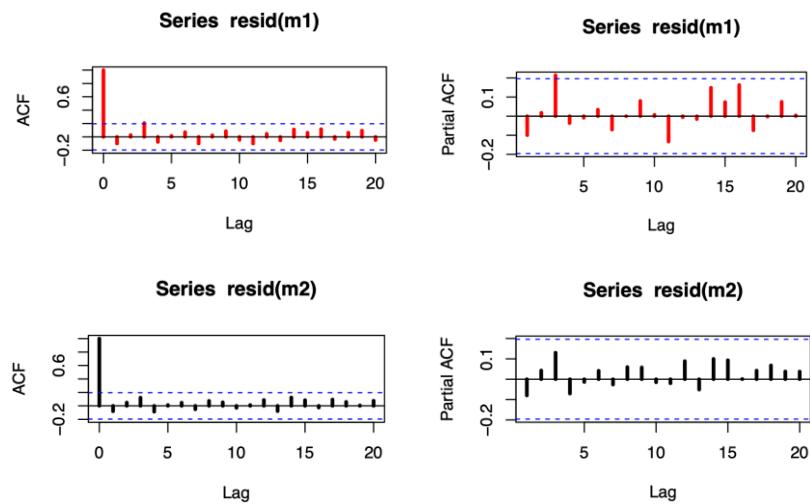
- Cuando existen este tipo de valores atípicos, se puede ver como la serie observada y la serie lineal teórica son iguales excepto por una observación en t_0



- Se puede ver como los residuos de la serie teórica linealizada estarán más en sintonía con los residuos de la serie en otros puntos en el tiempo, mientras que hay un desplazamiento de la curva de residuos para la serie original (sin incluir la variable exógena), ya que difiere durante momentos posteriores a la línea del otro modelo



- Si se comparan la ACF y la PACF de la serie observada con la del modelo lineal teórico, se puede ver como hay diferencias entre ambos modelos (por la inclusión de la función de transferencia en el modelo). Ahora, el modelo teórico permitirá detectar mejor cuál sería el modelo para la serie temporal sin perturbar



- Para el residuo con el mayor valor sobre un nivel especificado previamente, se puede hacer el siguiente procedimiento:
 - Primero se tienen que detectar todos los valores atípicos, basándose en un contraste de hipótesis para los tres tipos de valores atípicos

Obs	type_detected	W_coeff	ABS_L_Ratio
1	14	TC -0.21465793	5.997766
2	32	AO 0.19056486	6.298607
3	244	TC -0.15519774	4.735810
4	142	TC -0.12686443	3.960561
5	52	AO -0.09618053	3.417006
6	220	LS -0.10184648	3.285543
7	184	TC -0.11793107	3.889489
8	88	AO -0.08320794	3.114094
9	112	AO -0.08206789	3.115474
10	275	LS -0.08368511	2.874148
11	226	LS -0.08167645	2.838597
12	120	TC -0.08133733	2.861306

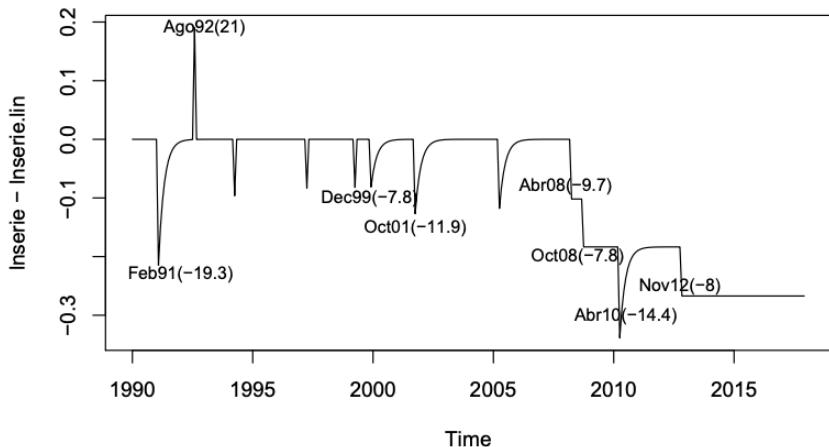
- Una manera de identificar los valores atípicos es ver si los residuos estandarizados son mayores a 2.9-3.4
- Los valores atípicos se pueden presentar a través de poner el tipo considerado, el coeficiente o peso que multiplicará a la función de transferencia en su regresión (la regresión β_i)
- El criterio que se usa se establece *a priori*, de modo que se establece un valor por el cuál una observación se considera un valor atípico
- Después, es necesario estimar el efecto de los valores atípicos más significativos o relevantes y estimar las series linealizadas para poder eliminar el valor atípico
 - Para ello, lo primero que es necesario es poder hacer una regresión lineal con la serie observada y restando las funciones de transferencia correspondientes para cada tipo de valor atípico

$$x_t = y_t - \sum_{i=1}^m \beta_i z_i$$

where z_i is transfer func.

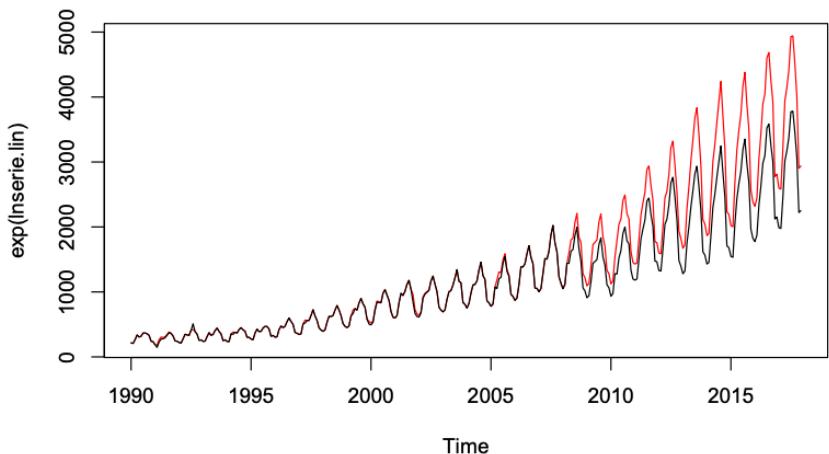
- En este caso, el efecto de los valores atípicos será estimado a partir de las β_i para $i = 1, 2, \dots, m$, siendo m el número de valores atípicos significativos. Una manera de analizar el efecto de los diferentes valores atípicos es a través de restar la serie linearizada a la serie original, dado que esto permite obtener los efectos:

$$y_t - x_t = \sum_{i=1}^m \beta_i z_i$$



- Aquellos que no sean lo suficientemente significativos se mantendrán en la muestra como si fueran observaciones comunes. Lo normal es tener pocos valores atípicos, dado que la presencia de muchos denota errores a la hora de estimar o de otros tipos
- Este proceso se tiene que repetir hasta que todos los residuos estén por encima de un límite o nivel concreto
 - Una vez se ha terminado el proceso, se puede hacer un gráfico comparando la serie observada originalmente con la serie linearizada o teórica teniendo en cuenta los valores atípicos
 - Si se ha aplicado una transformación logarítmica, se tiene que calcular la exponencial de esta serie linearizada para compararla con la original

y_t is transformed to $\ln(y_t)$ $\Rightarrow e^{x_t}$



- Un tipo de efectos externos importantes para considerar en el análisis de series temporales son los efectos de calendario para datos mensuales
 - Los efectos de calendario son aquellos efectos que, debido a la configuración del calendario, pueden afectar al fenómeno estudiado en la serie temporal
 - Normalmente, cada mes tiene el mismo número de días, excepto febrero en los años bisiestos
 - Las configuraciones normalmente se conocen porque son del calendario, de modo que, aunque sean futuras, se sabe cuándo y cómo ocurren
 - Hay algunos ejemplos muy claros y evidentes, tales como Semana Santa o los días comerciales o *trading days* en bolsa, aunque existen otros no tan comunes (como Navidad)
 - Los efectos de calendario como los de Semana Santa son efectos que ocurren porque la semana cae en unos meses diferentes dependiendo del año

marzo de 2016							abril de 2016						
L	M	X	J	V	S	D	L	M	X	J	V	S	D
			1	2	3	4	5	6		1	2	3	
7	8	9	10	11	12	13	4	5	6	7	8	9	10
14	15	16	17	18	19	20	11	12	13	14	15	16	17
21	22	23	24	25	26	27	18	19	20	21	22	23	24
28	29	30	31				25	26	27	28	29	30	



marzo de 2017							abril de 2017						
L	M	X	J	V	S	D	L	M	X	J	V	S	D
			1	2	3	4	5		1	2			
6	7	8	9	10	11	12	3	4	5	6	7	8	9
13	14	15	16	17	18	19	10	11	12	13	14	15	16
20	21	22	23	24	25	26	17	18	19	20	21	22	23
27	28	29	30	31			24	25	26	27	28	29	30

- Si la serie temporal se afecta por la Semana Santa, entonces las predicciones no tendrán en cuenta estos cambios (los efectos de calendario)
- Por lo tanto, para esos años en los que la Semana Santa está totalmente en marzo, se mueve la otra mitad del efecto a abril (se resta -0.5 del efecto a marzo y se suma 0.5 del efecto en marzo), y viceversa

marzo de 2016							abril de 2016						
L	M	X	J	V	S	D	L	M	X	J	V	S	D
1	2	3	4	5	6		1	2	3				
7	8	9	10	11	12	13	4	5	6	7	8	9	10
14	15	16	17	18	19	20	11	12	13	14	15	16	17
21	22	23	24	25	26	27	18	19	20	21	22	23	24
28	29	30	31				25	26	27	28	29	30	

→ Ideal
(March/April=0.5/0.5)

marzo de 2016							abril de 2016						
L	M	X	J	V	S	D	L	M	X	J	V	S	D
1	2	3	4	5	6		1	2					
7	8	9	10	11	12	13	3	4	5	6	7	8	9
13	14	15	16	17	18	19	10	11	12	13	14	15	16
20	21	22	23	24	25	26	17	18	19	20	21	22	23
27	28	29	30	31			24	25	26	27	28	29	30

marzo de 2017							abril de 2017						
L	M	X	J	V	S	D	L	M	X	J	V	S	D
1	2	3	4	5	6		1	2					
6	7	8	9	10	11	12	3	4	5	6	7	8	9
13	14	15	16	17	18	19	10	11	12	13	14	15	16
20	21	22	23	24	25	26	24	25	26	27	28	29	30
27	28	29	30	31			27	28	29	30	31		

- De este modo, se puede plantear una serie temporal auxiliar con las proporciones de días de Semana Santa que se quiere transferir al otro mes y cambiando su signo (dado que irán restando en la regresión). Esta serie auxiliar permitirá estimar el efecto de la Semana Santa en la serie temporal original observada

	Jan	Feb	Mar	Apr	May	Jun
1990	0.0000000	0.0000000	-0.5000000	0.5000000	0.0000000	0.0000000
1991	0.0000000	0.0000000	0.5000000	-0.5000000	0.0000000	0.0000000
1992	0.0000000	0.0000000	-0.5000000	0.5000000	0.0000000	0.0000000
1993	0.0000000	0.0000000	-0.5000000	0.5000000	0.0000000	0.0000000
1994	0.0000000	0.0000000	0.1666667	-0.1666667	0.0000000	0.0000000
1995	0.0000000	0.0000000	-0.5000000	0.5000000	0.0000000	0.0000000
1996	0.0000000	0.0000000	-0.5000000	0.5000000	0.0000000	0.0000000
1997	0.0000000	0.0000000	0.5000000	-0.5000000	0.0000000	0.0000000
1998	0.0000000	0.0000000	-0.5000000	0.5000000	0.0000000	0.0000000
1999	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000
2000	0.0000000	0.0000000	-0.5000000	0.5000000	0.0000000	0.0000000
2001	0.0000000	0.0000000	-0.5000000	0.5000000	0.0000000	0.0000000
2002	0.0000000	0.0000000	0.5000000	-0.5000000	0.0000000	0.0000000
2003	0.0000000	0.0000000	-0.5000000	0.5000000	0.0000000	0.0000000
2004	0.0000000	0.0000000	-0.5000000	0.5000000	0.0000000	0.0000000

- Los efectos de calendario como los de *trading days* son efectos que ocurren porque hay meses que tienen más días comerciables debido a la configuración de los fines de semana

octubre de 2011							octubre de 2012						
L	M	X	J	V	S	D	L	M	X	J	V	S	D
					1	2							
3	4	5	6	7	8	9							
10	11	12	13	14	15	16							
17	18	19	20	21	22	23							
24	25	26	27	28	29	30							
31													

- Si la serie temporal se afecta por *trading days*, entonces las predicciones no tendrán en cuenta estos cambios (los efectos de calendario)

- Cuando se mira un mes a lo largo de los años considerados, se puede ver como la proporción de *trading days* por fines de semana no siempre es constante (tendría que ser 5/2)
- Por lo tanto, para meses en donde faltan días de la semana, se suma el número de días necesarios para cumplir con la proporción 5/2, mientras que para meses en donde sobran días, se resta el número de días necesarios para cumplir con la proporción 5/2

octubre de 2011						octubre de 2012							
L	M	X	J	V	S	D	L	M	X	J	V	S	D
					1	2							
3	4	5	6	7	8	9	1	2	3	4	5	6	7
10	11	12	13	14	15	16	8	9	10	11	12	13	14
17	18	19	20	21	22	23	15	16	17	18	19	20	21
24	25	26	27	28	29	30	22	23	24	25	26	27	28
31							29	30	31				

Ideal
(TradingDays/WeekendDays=5/2)

octubre de 2011						octubre de 2012							
L	M	X	J	V	S	D	L	M	X	J	V	S	D
					1	2							
3	4	5	6	7	8	9	1	2	3	4	5	6	7
10	11	12	13	14	15	16	8	9	10	11	12	13	14
17	18	19	20	21	22	23	15	16	17	18	19	20	21
24	25	26	27	28	29	30	22	23	24	25	26	27	28
31	+	+	+	+			-29	-30	-31				

- De este modo, se puede plantear una serie temporal auxiliar con los días que se añaden o se restan para cada mes y cambiando su signo (dato que irán restando en la regresión). Esta serie auxiliar permitirá estimar el efecto de los *trading days* en la serie temporal original observada

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
1990	3.0	0.0	-0.5	-1.5	3.0	-1.5	-0.5	3.0	-5.0	3.0	2.0	-4.0
1991	3.0	0.0	-4.0	2.0	3.0	-5.0	3.0	-0.5	-1.5	3.0	-1.5	-0.5
1992	3.0	-2.5	-0.5	2.0	-4.0	2.0	3.0	-4.0	2.0	-0.5	-1.5	3.0
1993	-4.0	0.0	3.0	2.0	-4.0	2.0	-0.5	-0.5	2.0	-4.0	2.0	3.0
1994	-4.0	0.0	3.0	-1.5	-0.5	2.0	-4.0	3.0	2.0	-4.0	2.0	-0.5
1995	-0.5	0.0	3.0	-5.0	3.0	2.0	-4.0	3.0	-1.5	-0.5	2.0	-4.0
1996	3.0	1.0	-4.0	2.0	3.0	-5.0	3.0	-0.5	-1.5	3.0	-1.5	-0.5
1997	3.0	0.0	-4.0	2.0	-0.5	-1.5	3.0	-4.0	2.0	3.0	-5.0	3.0
1998	-0.5	0.0	-0.5	2.0	-4.0	2.0	3.0	-4.0	2.0	-0.5	-1.5	3.0
1999	-4.0	0.0	3.0	2.0	-4.0	2.0	-0.5	-0.5	2.0	-4.0	2.0	3.0
2000	-4.0	1.0	3.0	-5.0	3.0	2.0	-4.0	3.0	-1.5	-0.5	2.0	-4.0
2001	3.0	0.0	-0.5	-1.5	3.0	-1.5	-0.5	3.0	-5.0	3.0	2.0	-4.0
2002	3.0	0.0	-4.0	2.0	3.0	-5.0	3.0	-0.5	-1.5	3.0	-1.5	-0.5
2003	3.0	0.0	-4.0	2.0	-0.5	-1.5	3.0	-4.0	2.0	3.0	-5.0	3.0
2004	-0.5	-2.5	3.0	2.0	-4.0	2.0	-0.5	-0.5	2.0	-4.0	2.0	3.0

- De este modo, una vez se tienen las series temporales auxiliares para los efectos de calendario, se pueden usar como variables exógenas en la regresión linealizada (ARMAX) y así tener en cuenta el efecto

	wTradDays	wEast	serie	serieEC
Jan 2014	3.0	0.0000000	1429.618	1437.340
Feb 2014	0.0	0.0000000	1454.116	1454.116
Mar 2014	-4.0	-0.5000000	1865.363	1910.060
Apr 2014	2.0	0.5000000	2300.657	2238.765
May 2014	-0.5	0.0000000	2506.058	2503.809
Jun 2014	-1.5	0.0000000	2767.361	2759.917
Jul 2014	3.0	0.0000000	3010.583	3026.845
Aug 2014	-4.0	0.0000000	3246.128	3222.896
Sep 2014	2.0	0.0000000	2838.046	2848.257
Oct 2014	3.0	0.0000000	2461.470	2474.766
Nov 2014	-5.0	0.0000000	1709.285	1694.007
Dec 2014	3.0	0.0000000	1680.049	1689.124
Jan 2015	-0.5	0.0000000	1547.069	1545.681
Feb 2015	0.0	0.0000000	1533.699	1533.699
Mar 2015	-0.5	-0.1666667	2049.329	2068.662
Apr 2015	2.0	0.1666667	2437.182	2420.917
May 2015	-4.0	0.0000000	2670.680	2651.566
Jun 2015	2.0	0.0000000	2855.575	2865.849
Jul 2015	3.0	0.0000000	3190.774	3208.009
Aug 2015	-4.0	0.0000000	3352.763	3328.767
Sep 2015	2.0	0.0000000	2976.828	2987.538
Oct 2015	-0.5	0.0000000	2686.231	2683.820
Nov 2015	-1.5	0.0000000	1967.711	1962.418

- Este paso se tendría que hacer justo antes que el tratamiento de los valores atípicos, dado que, de otro modo, la estimación de los efectos de valores atípicos y de otros factores externos se verá sesgada
- Cuando se estima el modelo teniendo en cuenta los efectos de calendario como variables exógenas, es importante ver si el parámetro de estos efectos es significativo o no (con tal de saber si tienen un efecto significativo en la serie temporal)

Los elementos básicos de la teoría del dominio temporal

- Mucho del material desarrollado para la estimación de media cuadrada y la regresión se pueden incrustar en un marco más general que involucra un espacio de producto interior que también sea completo (que satisface la condición de Cauchy)
 - Un producto interior se denota como $\langle \mathbf{x}, \mathbf{y} \rangle$, y un espacio de producto interior se define como un conjunto de elementos \mathbf{x} y \mathbf{y} que cumplen con las siguientes propiedades para el producto interior:
 - (a) $\langle \mathbf{x}, \mathbf{y} \rangle = \langle \mathbf{y}, \mathbf{x} \rangle^*$
 - (b) $\langle \mathbf{x} + \mathbf{y}, \mathbf{z} \rangle = \langle \mathbf{x}, \mathbf{z} \rangle + \langle \mathbf{y}, \mathbf{z} \rangle$
 - (c) $\langle \alpha \mathbf{x}, \mathbf{y} \rangle = \alpha \langle \mathbf{x}, \mathbf{y} \rangle$
 - (d) $\langle \mathbf{x}, \mathbf{x} \rangle = \|\mathbf{x}\|^2 \geq 0$

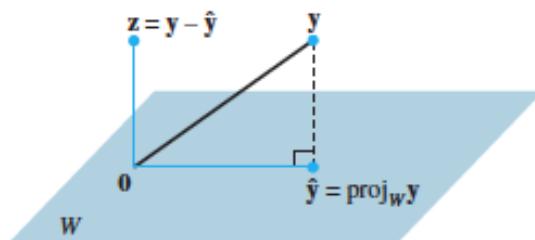
$$(e) \langle \mathbf{x}, \mathbf{x} \rangle = \mathbf{0} \text{ iff } \mathbf{x} = \mathbf{0}$$

- Dos ejemplos de productos interiores son $E(\mathbf{x}\mathbf{y}^*)$, donde los elementos son variables aleatorias, y $\sum x_i y_i^*$, donde los elementos son secuencias. Ambos ejemplos, igual que en las propiedades, incluyen la posibilidad de que haya elementos complejos, en donde “*” denota la conjugación
- La norma, por supuesto, satisface la desigualdad triangular y la desigualdad de Cauchy-Schwarz

$$\|\mathbf{x} + \mathbf{y}\|^2 \leq \|\mathbf{x}\|^2 + \|\mathbf{y}\|^2$$

$$|\langle \mathbf{x}, \mathbf{y} \rangle|^2 \leq \|\mathbf{x}\|^2 \|\mathbf{y}\|^2$$

- Un espacio de Hilbert \mathcal{H} se define como un espacio de producto interior con la propiedad de Cauchy, por lo que es un espacio de producto interior completo
 - Esto significa que todas las secuencias de Cauchy convergen en norma
- Para trabajar de manera general con series temporales, se enfatizan el teorema de la proyección y el principio de ortogonalidad asociado con tal de resolver problemas de estimación lineal
 - Siendo \mathcal{M} un subespacio cerrado del espacio de Hilbert \mathcal{H} y siendo \mathbf{y} un elemento en \mathcal{H} , \mathbf{y} solo puede ser representado como $\mathbf{y} = \hat{\mathbf{y}} + \mathbf{z}$, donde $\hat{\mathbf{y}}$ pertenece a \mathcal{M} y \mathbf{z} es ortogonal a \mathcal{M} (de modo que $\langle \mathbf{z}, \mathbf{w} \rangle = 0$ para toda \mathbf{w} en \mathcal{M})



- Además, el punto $\hat{\mathbf{y}}$ es el más cercano a \mathbf{y} en el sentido de que, para cualquier \mathbf{w} en \mathcal{M} , $\|\mathbf{y} - \mathbf{w}\| \geq \|\mathbf{y} - \hat{\mathbf{y}}\|$ donde la igualdad solo ocurre cuando $\mathbf{w} = \hat{\mathbf{y}}$
- Todo esto permite obtener la propiedad de la ortogonalidad, la cual expresa que para toda $\mathbf{w} \in \mathcal{M}$, se cumple que $\langle \mathbf{z}, \mathbf{w} \rangle =$

$\langle \mathbf{y} - \hat{\mathbf{y}}, \mathbf{w} \rangle = 0$. Esto se puede usar para encontrar fácilmente la expresión de la proyección

- La norma del error se puede desarrollar de la siguiente manera debido a la propiedad de la ortogonalidad entre \mathbf{z} y $\mathbf{w} \in \mathcal{M}$:

$$\begin{aligned}\|\mathbf{y} - \hat{\mathbf{y}}\|^2 &= \langle \mathbf{y} - \hat{\mathbf{y}}, \mathbf{y} - \hat{\mathbf{y}} \rangle = \langle \mathbf{y} - \hat{\mathbf{y}}, \mathbf{y} \rangle - \langle \mathbf{y} - \hat{\mathbf{y}}, \hat{\mathbf{y}} \rangle = \\ &= \langle \mathbf{y} - \hat{\mathbf{y}}, \mathbf{y} \rangle - \langle \mathbf{z}, \hat{\mathbf{y}} \rangle = \langle \mathbf{y} - \hat{\mathbf{y}}, \mathbf{y} \rangle = \langle \mathbf{z}, \mathbf{y} \rangle \\ &\text{as } \langle \mathbf{z}, \hat{\mathbf{y}} \rangle = 0 \text{ due to } \hat{\mathbf{y}} \in \mathcal{M}\end{aligned}$$

- El mapeado $P_{\mathcal{M}}\mathbf{y} = \hat{\mathbf{y}}$ para $\mathbf{y} \in \mathcal{H}$ se denomina el mapeado de proyección de \mathcal{H} sobre \mathcal{M} . Además, el *span* cerrado de un conjunto finito $\{\mathbf{x}_1, \dots, \mathbf{x}_n\}$ de elementos del espacio de Hilbert \mathcal{H} , denotado como $\mathcal{M} = \text{span}(\{\mathbf{x}_1, \dots, \mathbf{x}_n\})$ se define como todas las combinaciones lineales $\mathbf{w} = a_1\mathbf{x}_1 + \dots + a_n\mathbf{x}_n$, donde a_1, \dots, a_n son escalares
- Por el teorema de proyección, la proyección de $\mathbf{y} \in \mathcal{H}$ sobre \mathcal{M} es única y se da por la siguiente expresión:

$$P_{\mathcal{M}}\mathbf{y} = \hat{\mathbf{y}} = a_1\mathbf{x}_1 + \dots + a_n\mathbf{x}_n$$

- En este caso, el conjunto de escalares $\{a_1, \dots, a_n\}$ se encuentran utilizando el principio de ortogonalidad:

$$\langle \mathbf{y} - P_{\mathcal{M}}\mathbf{y}, \mathbf{x}_j \rangle = \langle \mathbf{y}, \mathbf{x}_j \rangle - \langle P_{\mathcal{M}}\mathbf{y}, \mathbf{x}_j \rangle = 0 \quad \text{for } j = 1, \dots, n$$

- Evidentemente, $\{a_1, \dots, a_n\}$ se puede obtener resolviendo la siguiente ecuación (derivada de la anterior):

$$\langle P_{\mathcal{M}}\mathbf{y}, \mathbf{x}_j \rangle = \sum_{i=1}^n a_i \langle \mathbf{x}_i, \mathbf{x}_j \rangle \Rightarrow \langle \mathbf{y}, \mathbf{x}_j \rangle = \sum_{i=1}^n a_i \langle \mathbf{x}_i, \mathbf{x}_j \rangle$$

- Algunos resultados útiles pertenecientes a los mapeados de proyecciones son los siguientes:

- (a) $P_{\mathcal{M}}(a\mathbf{x} + b\mathbf{y}) = aP_{\mathcal{M}}\mathbf{x} + bP_{\mathcal{M}}\mathbf{y}$ for $\mathbf{x}, \mathbf{y} \in \mathcal{H}$
- (b) If $\|\mathbf{y}_n - \mathbf{y}\| \rightarrow 0$, then $P_{\mathcal{M}}\mathbf{y}_n \rightarrow P_{\mathcal{M}}\mathbf{y}$ as $n \rightarrow \infty$
- (c) $\mathbf{w} \in \mathcal{M}$ iff $P_{\mathcal{M}}\mathbf{w} = \mathbf{w}$, so that $P_{\mathcal{M}_1}(P_{\mathcal{M}_2}\mathbf{y}) = P_{\mathcal{M}_1}\mathbf{y}$ for $\forall \mathbf{y} \in \mathcal{H}$
- (d) Being \mathcal{M}_1 and \mathcal{M}_2 closed subspaces of \mathcal{H} , $\mathcal{M}_1 \subseteq \mathcal{M}_2$ iff $P_{\mathcal{M}_1}(P_{\mathcal{M}_2}\mathbf{y}) = P_{\mathcal{M}_1}\mathbf{y}$ for $\forall \mathbf{y} \in \mathcal{H}$
- (e) Being \mathcal{M} a closed subspace of \mathcal{H} & \mathcal{M}_{\perp} the orthog. complem.

of $\mathcal{M}, \mathcal{M}_\perp$ is also a closed subspace of \mathcal{H} , and for $\forall \mathbf{y} \in \mathcal{H}$,

$$\mathbf{y} = P_{\mathcal{M}}\mathbf{y} + P_{\mathcal{M}_\perp}\mathbf{y}$$

- La parte (c) lleva a un conocido resultado de los modelos lineales, en donde una matriz cuadrada \mathbf{M} es una matriz de proyección si, y solo si, es simétrica e idempotente ($\mathbf{M}^2 = \mathbf{M}$). En el caso de la regresión lineal sería $\mathbf{M} = \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'$
- Las partes (d) y (e) del teorema son útiles para establecer soluciones recursivas para la estimación y la predicción
- Cuando los elementos de \mathcal{H} son vectores, el problema planteado es el problema de la regresión lineal
 - Considerando los vectores $\mathbf{y} = (y_1, \dots, y_n)'$ y $\mathbf{z}_i = (z_{1i}, \dots, z_{ni})'$ para $i = 1, \dots, q$, se pueden obtener los siguientes resultados para la estimación de la proyección del vector \mathbf{y} observado en el *span* de $\beta_1\mathbf{z}_1 + \dots + \beta_q\mathbf{z}_q$:
 - A partir de la condición de ortogonalidad anterior, es posible obtener los siguientes resultados:

$$\begin{aligned} \langle \mathbf{z}_i, \mathbf{y} \rangle &= \sum_{t=1}^n z_{ti} y_t = \mathbf{z}_i' \mathbf{y} \\ \Rightarrow \langle \mathbf{y} - \sum_{t=1}^q \beta_i \mathbf{z}_i, \mathbf{z}_j \rangle &= \langle \mathbf{y}, \mathbf{z}_j \rangle - \langle \sum_{t=1}^q \beta_i \mathbf{z}_i, \mathbf{z}_j \rangle = \\ &= \langle \mathbf{y}, \mathbf{z}_j \rangle - \sum_{t=1}^q \beta_i \langle \mathbf{z}_i, \mathbf{z}_j \rangle = 0 \\ \Rightarrow \mathbf{y}' \mathbf{z}_j &= \sum_{t=1}^q \beta_i \mathbf{z}_i' \mathbf{z}_j \quad \text{for } j = 1, 2, \dots, q \end{aligned}$$

- Usando notación matricial, se considera una matriz $\mathbf{Z} = (\mathbf{z}_1, \dots, \mathbf{z}_q)$ y un vector $\boldsymbol{\beta} = (\beta_1, \dots, \beta_n)$ se asume que tiene rango completo para poder obtener los siguientes resultados:

$$\mathbf{y}' \mathbf{Z} = \boldsymbol{\beta}' (\mathbf{Z}' \mathbf{Z}) \Rightarrow \hat{\boldsymbol{\beta}} = (\mathbf{Z}' \mathbf{Z})^{-1} \mathbf{Z}' \mathbf{y}$$

- El error medio cuadrático, en este caso sería el siguiente:

$$\begin{aligned} \left\| \mathbf{y} - \sum_{i=1}^q \hat{\beta}_i \mathbf{z}_i \right\|^2 &= \langle \mathbf{y} - \sum_{i=1}^q \hat{\beta}_i \mathbf{z}_i, \mathbf{y} \rangle = \langle \mathbf{y}, \mathbf{y} \rangle - \sum_{i=1}^q \hat{\beta}_i \langle \mathbf{z}_i, \mathbf{y} \rangle = \\ &= \mathbf{y}' \mathbf{y} - \hat{\boldsymbol{\beta}}' \mathbf{Z}' \mathbf{y} \end{aligned}$$

- Aunque la generalidad parezca innecesaria para el caso finito, es conveniente pensar en los elementos de \mathcal{H} con dimensionalidad infinita, de modo que el principio de ortogonalidad es útil
 - Por ejemplo, para la proyección del proceso $\{\mathbf{x}_t : t = 0, \pm 1, \pm 2, \dots\}$ en la variedad (matemática) lineal abarcada (en el sentido matemático) por todas las convoluciones de la forma $\hat{\mathbf{x}} = \sum_{k=-\infty}^{\infty} a_k \mathbf{x}_{t-k}$ serían de esta forma
 - De este modo, el conjunto de escalares que multiplican los vectores sería infinito
- Añadiendo estructura extra, la esperanza condicional se puede definir como un mapeado de proyección para variables aleatorias en L^2 , de modo que la esperanza condicional debe satisfacer el teorema de la proyección, el principio de ortogonalidad y los resultados derivados
 - Se puede definir la esperanza condicional $E_{\mathcal{M}}$ como mapeado de variables aleatorias en L^2 con la relación de equivalencia de que, para $\mathbf{x}, \mathbf{y} \in L^2$, $\mathbf{x} = \mathbf{y}$ si $P(\mathbf{x} = \mathbf{y}) = 1$
 - En particular, para $\mathbf{y} \in L^2$, si \mathcal{M} es un subespacio cerrado de L^2 conteniendo 1, la esperanza condicional de \mathbf{y} dado \mathcal{M} se define como la proyección de \mathbf{y} en \mathcal{M} , por lo que $E_{\mathcal{M}}(\mathbf{y}) = P_{\mathcal{M}}(\mathbf{y})$
 - Si $\mathcal{M}(\mathbf{x})$ denota el subespacio cerrado de todas las variables aleatorias en L^2 que se puede escribir como funciones medibles de \mathbf{x} , entonces se puede definir para $\mathbf{x}, \mathbf{y} \in L^2$, la esperanza condicional de \mathbf{y} dado \mathbf{x} como $E(\mathbf{y}|\mathbf{x}) = E_{\mathcal{M}(\mathbf{x})}(\mathbf{y})$
 - Esta idea se puede generalizar de manera obvia para definir la esperanza condicional de \mathbf{y} dado $\mathbf{x}_{1:n} = (\mathbf{x}_1, \dots, \mathbf{x}_n)$, de modo que $E(\mathbf{y}|\mathbf{x}_{1:n}) = E_{\mathcal{M}(\mathbf{x}_{1:n})}(\mathbf{y})$
 - Bajo la notación y las condiciones anteriores, si $(\mathbf{x}_1, \dots, \mathbf{x}_n)$ es normal multivariante, entonces se cumple la siguiente igualdad:

$$E(\mathbf{y}|\mathbf{x}_{1:n}) = P_{\text{span}(\{\mathbf{x}_1, \dots, \mathbf{x}_n\})} \mathbf{y}$$

- Debido al teorema de proyección, $E(\mathbf{y}|\mathbf{x}_{1:n})$ es el único elemento $E_{\mathcal{M}(\mathbf{x})}(\mathbf{y})$ que satisface el principio de ortogonalidad, y se tiene que demostrar que $E_{\mathcal{M}(\mathbf{x})}(\mathbf{y}) = P_{\text{span}(\{\mathbf{x}_1, \dots, \mathbf{x}_n\})}\mathbf{y}$

$$E[(\mathbf{y} - E_{\mathcal{M}(\mathbf{x})}(\mathbf{y}))\mathbf{w}] = 0 \quad \text{for } \forall \mathbf{w} \in \mathcal{M}(\mathbf{x})$$

- El teorema de la proyección hace que $\hat{\mathbf{y}}$ satisfaga $\langle \mathbf{y} - \hat{\mathbf{y}}, \mathbf{x}_i \rangle = 0$ para $i = 0, 1, \dots, n$ (donde $\mathbf{x}_0 = \mathbf{1}$). No obstante, como el producto interior equivale a la covarianza y el vector de variables $(\mathbf{y} - \hat{\mathbf{y}}, \mathbf{x}_1, \dots, \mathbf{x}_n)$ es normal multivariante, entonces ambos términos son independientes

$$\langle \mathbf{y} - \hat{\mathbf{y}}, \mathbf{x}_i \rangle = E[(\mathbf{y} - \hat{\mathbf{y}})\mathbf{x}_i] = \text{Cov}(\mathbf{y} - \hat{\mathbf{y}}, \mathbf{x}_i) = 0$$

$\Rightarrow \mathbf{y} - \hat{\mathbf{y}} \text{ & } \mathbf{x}_i \text{ independent due to Gaussian dist.}$

- Por lo tanto, si $\mathbf{w} \in \mathcal{M}(\mathbf{x}) = \text{span}(\{\mathbf{x}_1, \dots, \mathbf{x}_n\})$, entonces \mathbf{w} y $\mathbf{y} - \hat{\mathbf{y}}$ son independientes y, por tanto, $\langle \mathbf{y} - \hat{\mathbf{y}}, \mathbf{w} \rangle = E[(\mathbf{y} - \hat{\mathbf{y}})\mathbf{w}] = E[(\mathbf{y} - \hat{\mathbf{y}})]E(\mathbf{w}) = 0$ (porque $\langle \mathbf{y} - \hat{\mathbf{y}}, \mathbf{1} \rangle = E[(\mathbf{y} - \hat{\mathbf{y}})] = 0$). Esto quiere decir que, en el caso normal, la esperanza condicional y la predicción lineal son equivalentes
- En el caso normal, la esperanza condicional tiene una forma explícita. Tomando dos vectores $\mathbf{y} = (y_1, \dots, y_n)'$ y $\mathbf{x} = (x_1, \dots, x_n)'$ y suponiendo que son conjuntamente normales, se obtienen las siguientes expresiones:

$$\begin{pmatrix} \mathbf{y} \\ \mathbf{x} \end{pmatrix} \sim N_{m+n} \left[\begin{pmatrix} \boldsymbol{\mu}_y \\ \boldsymbol{\mu}_x \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma}_{yy} & \boldsymbol{\Sigma}_{yx} \\ \boldsymbol{\Sigma}_{xy} & \boldsymbol{\Sigma}_{xx} \end{pmatrix} \right]$$

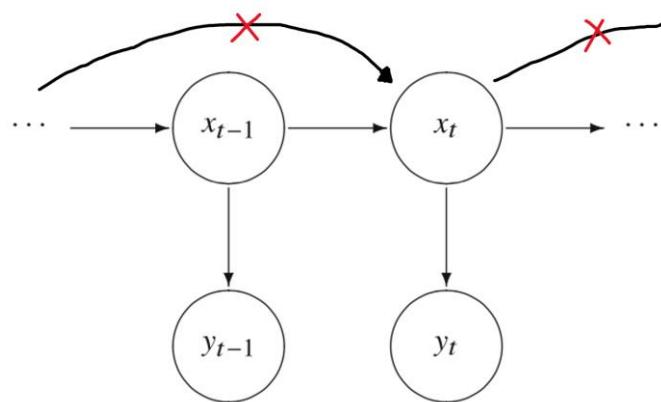
$$\Rightarrow \mathbf{y}|\mathbf{x} \sim N(\boldsymbol{\mu}_{y|x}, \boldsymbol{\Sigma}_{y|x}) \text{ where } \begin{cases} \boldsymbol{\mu}_{y|x} = \boldsymbol{\mu}_y + \boldsymbol{\Sigma}_{yx}\boldsymbol{\Sigma}_{xx}^{-1}(\mathbf{x} - \boldsymbol{\mu}_x) \\ \boldsymbol{\Sigma}_{y|x} = \boldsymbol{\Sigma}_{yy} - \boldsymbol{\Sigma}_{yx}\boldsymbol{\Sigma}_{xx}^{-1}\boldsymbol{\Sigma}_{xy} \end{cases}$$

- En este caso, se asume que $\boldsymbol{\Sigma}_{xx}$ es una matriz invertible o no singular

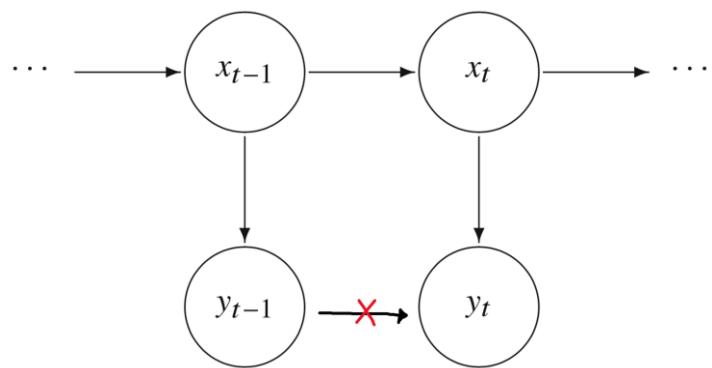
Los modelos de estado-espacio: el modelo lineal normal

- Un modelo muy general que engloba una clase entera de casos especiales de interés en el mismo modo que hace una regresión lineal es el modelo de estado-espacio o modelo lineal dinámico
 - El modelo nació en un contexto de seguimiento espacial, en donde la ecuación de estado define las ecuaciones de moción para la posición o estado de la nave espacial con localización \mathbf{x}_t y los datos \mathbf{y}_t reflejan la información que se puede observar del dispositivo de seguimiento

- Aunque este modelo nació para la investigación espacial, estos modelos se han aplicado a las áreas de economía, medicina y otras ciencias
 - En general, el modelo de estado espacio se caracteriza por dos principios importantes, las cuales son las siguientes:
 - La primera condición es que hay un proceso latente o escondido x_t , llamado proceso de estado. Este proceso de estado se asume que es un proceso de Markov, por lo que el futuro $\{x_s : s > t\}$ y el pasado $\{x_s : s < t\}$ son independientes si se condiciona al presente x_t



- La segunda condición es que las observaciones de y_t son independientes dados los estados de x_t , de modo que la dependencia entre observaciones de y_t se genera por los estados, pero no entre ellas



- El modelo de estado-espacio más utilizado y más simple es el gaussiano, del cual se pueden derivar varios casos especiales útiles
 - El modelo de estado-espacio lineal gaussiano, en su forma básica, emplea un modelo una autorregresión vectorial p -dimensional de primer orden como ecuación de estado

$$\mathbf{x}_t = \Phi \mathbf{x}_{t-1} + \mathbf{w}_t$$

- En este caso, el término \mathbf{x}_t es un vector de tamaño $p \times 1$ cuyos componentes y Φ es una matriz $p \times p$. Además, el término \mathbf{w}_t es un vector normal de tamaño $p \times 1$ cuyos componentes son independientes e idénticamente distribuidos con matriz de varianzas y covarianzas \mathbf{Q}

$$\mathbf{w}_t \sim N_p(\mathbf{0}, \mathbf{Q})$$

- En este modelo, se asume que el proceso comienza en un vector normal \mathbf{x}_0 , de modo que $\mathbf{x}_0 \sim N_p(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$
- No se observa el vector de estado \mathbf{x}_t directamente, sino que se observa una versión lineal transformada de este sumando un ruido, de modo que se obtiene la ecuación de observación:

$$\mathbf{y}_t = \mathbf{A}_t \mathbf{x}_t + \mathbf{v}_t$$

- En este caso, \mathbf{y}_t es un vector $q \times 1$ (de una dimensión mayor o menor a la de \mathbf{x}_t) y \mathbf{A}_t es una matriz de tamaño $p \times q$ llamada matriz de observación o de medida
 - El ruido aditivo de las observaciones es un vector independiente e idénticamente distribuido $\mathbf{v}_t \sim N_q(\mathbf{0}, \mathbf{R})$
 - Además, se asume inicialmente (por simplicidad) que $\mathbf{x}_0, \{\mathbf{w}_t\}$ y $\{\mathbf{v}_t\}$ no tienen correlación. Esta suposición no es necesaria, pero permite una mejor explicación de los conceptos básicos
 - Por lo tanto, el modelo se puede desarrollar de manera explícita para mostrar los parámetros y variables involucradas
 - Aunque el modelo parezca simplista, es muy general. Si el proceso de estado es una autorregresión vectorial de orden s , el modelo se puede expresar como un modelo s -dimensional:
- $$\begin{pmatrix} \mathbf{x}_t \\ \mathbf{x}_{t-1} \end{pmatrix} = \begin{pmatrix} \Phi_1 & \Phi_2 \\ I & \mathbf{0} \end{pmatrix} \begin{pmatrix} \mathbf{x}_{t-1} \\ \mathbf{x}_{t-2} \end{pmatrix} + \begin{pmatrix} \mathbf{w}_t \\ \mathbf{0} \end{pmatrix}$$
- $$\mathbf{y}_t = [\mathbf{A}_t | \mathbf{0}] \begin{pmatrix} \mathbf{x}_t \\ \mathbf{x}_{t-1} \end{pmatrix} + \mathbf{v}_t$$
- En este caso, el vector de \mathbf{x} tiene tamaño $2p \times 1$ y $[\mathbf{A}_t | \mathbf{0}]$ es una matriz adjunta de tamaño $q \times 2p$ en la que \mathbf{A}_t es la matriz de observación

- Las ventajas reales de la formulación de estado espacio provienen de su formulación matricial
 - Las formas especiales que se pueden desarrollar para varias versiones de la matriz A_t y para el esquema de transiciones definido por la matriz Φ permiten ajustar estructuras más parsimoniosas con menos parámetros necesarios para describir una serie temporal multivariante
 - Existen muchos ejemplos, y los modelos estructurales son un buen ejemplo de la flexibilidad de estos modelos
- Igual que en el modelo ARIMAX, variables exógenas o insumos fijos pueden entrar en la ecuación de estado o de observación
 - En este caso, se supone que se tiene un vector u_t de tamaño $r \times 1$ y se escribe un modelo como el siguiente:
$$x_t = \Phi x_{t-1} + \Upsilon u_t + w_t$$

$$y_t = A_t x_t + \Gamma u_t + v_t$$
 - En este caso, Υ es una matriz $p \times r$ y Γ es una matriz $q \times r$, aunque cualquiera de las dos matrices puede ser una matriz de ceros (con tal de eliminarla de una u otra ecuación)
- La introducción del enfoque de estado-espacio como una herramienta para modelar los datos en las ciencias sociales y biomédicas requieren la identificación del modelo y la estimación de los parámetros, dado que rara vez se tiene una ecuación diferencial bien definida describiendo la transición del estado
 - Las cuestiones de interés general para el modelo dinámico lineal tienen relación con estimar $\Phi, \Upsilon, A_t, \Gamma, Q$ y R , las cuales definen el modelo particular, y estimar o predecir valores para el proceso no observado x_t subyacente
 - Las ventajas de la formulación del modelo de estado-espacio están en la facilidad con la que se pueden tratar varias configuraciones de datos perdidos y la cantidad de modelos que se pueden generar
 - La analogía entre la matriz de observaciones A_t y la matriz de diseño en una regresión y análisis de varianza es útil. Se pueden generar estructuras de efectos fijos y aleatorios que son constantes o varían en el tiempo simplemente al escoger apropiadamente la matriz A_t y la matriz Φ

- Es instructivo considerar un modelo univariante simple en donde se observa un proceso $AR(1)$ usando un instrumento con ruido, de modo que la ecuación de estado es x_t y la observada y_t es una función de x_t

$$x_t = \phi x_{t-1} + w_t$$

$$y_t = x_t + v_t$$

- En este caso, se considera que los errores se distribuyen $w_t \sim iid N(0, \sigma_w^2)$ y $v_t \sim iid N(0, \sigma_v^2)$, y que $x_0 \sim N\left(0, \frac{\sigma_w^2}{1-\phi^2}\right)$. Además, se asume que $\{w_t\}$, $\{v_t\}$ y x_0 son independientes y $t = 1, 2, \dots$
- Aunque se haya estudiado el comportamiento de un proceso $AR(1)$ anteriormente, ahora se tiene que investigar el efecto del ruido en sus dinámicas. Debido a que se asume que x_t es estacionaria, entonces y_t también lo es, y se pueden obtener las siguientes funciones de autocovarianza y autocorrelación:

$$\gamma_y(0) = Var(x_t + v_t) = \frac{\sigma_w^2}{1 - \phi^2} + \sigma_v^2$$

$$\gamma_y(h) = Cov(x_t + v_t, x_{t-h} + v_{t-h}) = \gamma_x(h) \text{ for } h \geq 1$$

$$\Rightarrow \rho_y(h) = \frac{\gamma_y(h)}{\gamma_y(0)} = \left(1 + \frac{\sigma_v^2}{\sigma_w^2}(1 - \phi^2)\right)^{-1} \phi^h$$

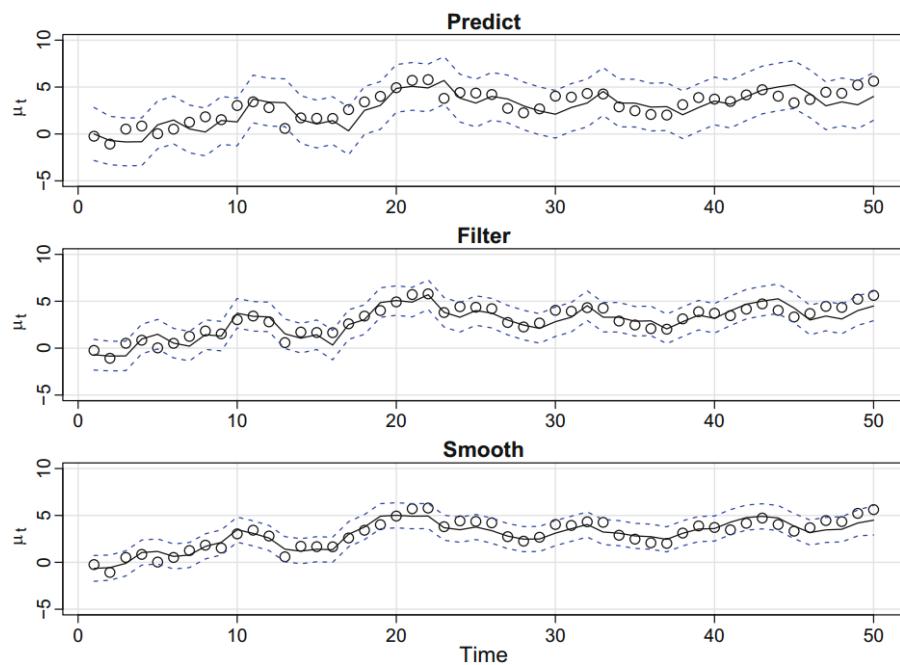
- Como se puede ver en la correlación, las observaciones y_t no son $AR(1)$ a menos que $\sigma_v^2 = 0$. Además, la estructura de autocorrelación de y_t es idéntica a la estructura de autocorrelación de un $ARMA(1,1)$. Por lo tanto, las observaciones se pueden reescribir en esta forma, escogiendo unas θ y σ_u^2 escogidas adecuadamente:

$$y_t = \phi y_{t-1} + \theta u_{t-1} + u_t \text{ where } u_t \sim iid N(0, \sigma_u^2)$$

- Aunque haya una equivalencia entre modelos ARMA estacionarios y modelos de estado-espacio estacionarios, a veces es más fácil trabajar con una forma u otra. En caso de que haya datos perdidos, efectos mixtos, o ciertos tipos de no estacionariedad, es más fácil trabajar con la última clase de modelos

Los modelos de estado-espacio: Kalman y estimación

- Desde un punto de vista práctico, un objetivo primario de cualquier análisis involucrando el modelo de estado-espacio sería producir estimadores para la señal subyacente no observada x_t dados unos datos $y_{1:s} = \{y_1, y_2, \dots, y_s\}$ hasta el momento s
 - Como se verá, la del estado es un componente esencial para la estimación de los parámetros, por lo que se tienen que usar los conceptos de predicción, filtro y suavizado



- Cuando $s < t$, el problema es de predicción o *forecasting*; cuando $s = t$, el problema es de filtro o *filter*; y cuando $s > t$, el problema es de suavizado o *smoothing*
- Además de estimaciones, se querría medir la precisión de estas, por lo que se utilizará el filtro y el suavizante de Kalman
- Durante esta sección, se utilizará la notación x_t^s para la esperanza condicional de x_t a los datos y la matriz P_{t_1,t_2}^s para la matriz de autocovarianzas para dos momentos t_1 y t_2

$$x_t^s = E(x_t | y_{1:s})$$

$$P_{t_1,t_2}^s = E \left[(x_{t_1} - x_{t_1}^s)(x_{t_2} - x_{t_2}^s)' \right]$$
 - Cuando $t_1 = t_2 = t$, se utiliza P_t^s por conveniencia notacional

- Al obtener las ecuaciones de filtración y suavizado, uno se apoyará en la suposición de normalidad, pero, aunque no se esté en el caso gaussiano, los estimadores que se obtendrán son los que minimizan el error cuadrático dentro de la clase de los estimadores lineales
 - Por lo tanto, se puede entender la esperanza condicional como un operador de proyección (como en una regresión lineal) más que como una media ponderada y también se puede entender $\mathbf{y}_{1:s}$ como el espacio de combinaciones lineales de $\{\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_s\}$
 - En este caso, \mathbf{P}_{t_1,t_2}^s corresponde al error medio cuadrático, y como los procesos se asumen gaussianos, \mathbf{P}_{t_1,t_2}^s también se puede interpretar como la covarianza condicional de los errores
- $$\mathbf{P}_{t_1,t_2}^s = E \left[(\mathbf{x}_{t_1} - \mathbf{x}_{t_1}^s)(\mathbf{x}_{t_2} - \mathbf{x}_{t_2}^s)' \mid \mathbf{y}_{1:s} \right]$$
- Esto se puede ver a través de ver que la matriz de varianzas y covarianzas de $(\mathbf{x}_t - \mathbf{x}_t^s)$ e $\mathbf{y}_{1:s}$ es la matriz $\mathbf{0}$ para cualquier t y s , por lo que se puede ver que son ortogonales. Esto implica que ambos elementos son independientes (al ser gaussianas y no correlacionadas) y, por tanto, la distribución condicional de $(\mathbf{x}_t - \mathbf{x}_t^s)$ dado $\mathbf{y}_{1:s}$ es la distribución incondicional de $(\mathbf{x}_t - \mathbf{x}_t^s)$
- Para el modelo de estado-espacio especificado anteriormente, con condiciones iniciales $\mathbf{x}_0 = \boldsymbol{\mu}_0$ y $\mathbf{P}_0^0 = \boldsymbol{\Sigma}_0$ para $t = 1, 2, \dots, n$, se obtienen las siguientes igualdades

$$\mathbf{x}_t^{t-1} = \Phi \mathbf{x}_{t-1}^{t-1} + \boldsymbol{\Upsilon} \mathbf{u}_t \quad \text{with} \quad \mathbf{x}_t^t = \mathbf{x}_t^{t-1} + \mathbf{K}_t (\mathbf{y}_t - \mathbf{A}_t \mathbf{x}_t^{t-1} - \boldsymbol{\Gamma} \mathbf{u}_t)$$

$$\mathbf{P}_t^{t-1} = \Phi \mathbf{P}_{t-1}^{t-1} \Phi' + \mathbf{Q} \quad \text{with} \quad \mathbf{P}_t^t = (\mathbf{I} - \mathbf{K}_t \mathbf{A}_t) \mathbf{P}_t^{t-1}$$
 - Las ecuaciones que permiten actualizar las observaciones para \mathbf{x} y para \mathbf{P} son las ecuaciones de innovación (a la derecha de las del modelo), y estas se pueden interpretar como las ecuaciones de un filtro para cada elemento (ya que $s = t$)
 - La predicción para $t > n$, en cambio, se obtiene a través de las dos ecuaciones cuando las condiciones iniciales son \mathbf{x}_n^n y \mathbf{P}_n^n (que serían ecuaciones para el filtro de observaciones pasadas)
 - En este caso, \mathbf{K}_t es la ganancia de Kalman, la cual se puede interpretar como una cantidad que controla la magnitud en la que se confía en la estimación sobre las medidas observadas (controla el *trade-off*). Esta se define de la siguiente manera:

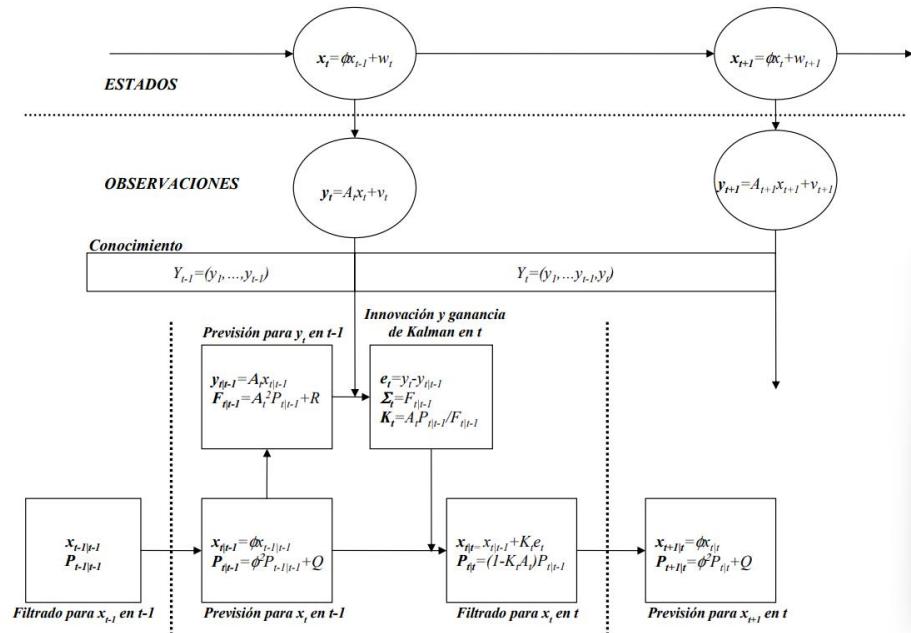
$$K_t = P_t^{t-1} A'_t [A_t P_t^{t-1} A'_t + R]^{-1}$$

- Los resultados importantes obtenidos a partir de las predicciones son las innovaciones (errores de predicción) y las matrices de varianzas y covarianzas correspondientes para $t = 1, 2, \dots, n$. Se asume que $\Sigma_t > 0$, lo cual se garantiza si $R > 0$ (aunque esta suposición no es necesaria y se podría relajar)

$$\epsilon_t = y_t - E_t(y_t | y_{1:t-1}) = y_t - A_t x_t^{t-1} - \Gamma u_t$$

$$\Sigma_t = Cov(\epsilon_t) = Cov[A_t(x_t - x_t^{t-1}) + v_t] = A_t P_t^{t-1} A'_t + R$$

- A partir de estas ecuaciones, es posible obtener un algoritmo para poder hacer una predicción de las observaciones a través de los estados y obtener un filtro en cada momento para los datos. A este proceso se le denomina filtro de Kalman (debido a las ecuaciones de filtro)



- Asumiendo que se comienza con un filtro (o con un valor inicial) en $t-1$, primero se calcula la predicción para x_t (x_t^{t-1}) y para y_t (y_t^{t-1}), además de sus varianzas
- Después, se obtiene el cálculo de la ganancia de Kalman y de las innovaciones para las observaciones, y usando esta información, se calcula el filtro de las observaciones para t (x_t^t y P_t^t)

Prediction of x	Prediction of y	Observation	Filter of x
$x_{1\parallel 0}, P_{1\parallel 0}$	$y_{1\parallel 0}, F_{1\parallel 0}$	y_1	$x_{0\parallel 0}, P_{0\parallel 0}$
$x_{2\parallel 1}, P_{2\parallel 1}$	$y_{2\parallel 1}, F_{2\parallel 1}$	y_2	$x_{1\parallel 1}, P_{1\parallel 1}$
$x_{3\parallel 2}, P_{3\parallel 2}$	$y_{3\parallel 2}, F_{3\parallel 2}$	y_3	$x_{2\parallel 2}, P_{2\parallel 2}$
:	:	:	$x_{3\parallel 3}, P_{3\parallel 3}$

- La demostración de los resultados anteriores no incluye la posibilidad en la que algunos o todos los parámetros varían en el tiempo, o la posibilidad de que las dimensiones de las observaciones cambien en el tiempo
 - No obstante, se puede obtener un corolario del filtro de Kalman para esta situación
 - Si en las ecuaciones del modelo uno o más parámetros dependen del tiempo ($\Phi = \Phi_t$, $\Upsilon = \Upsilon_t$ o $Q = Q_t$ en la ecuación de estado, o $\Gamma = \Gamma_t$ o $R = R_t$ en la ecuación de las observaciones), o las dimensiones de la ecuación de las observaciones depende del tiempo ($q = q_t$), los resultados anteriormente vistos se mantienen con las sustituciones apropiadas
- Es posible explorar el modelo, el filtro y la predicción desde el punto de vista de la densidad de probabilidad. En el modelo de estado-espacio visto, siendo p_Θ una función de densidad genérica paramétrica, el proceso cumple con la propiedad de Markov y las observaciones son condicionalmente independientes dados los estados

$$p_\Theta(x_t | x_{t-1}, \dots, x_1, x_0) = p_\Theta(x_t | x_{t-1})$$

$$p_\Theta(y_{1:n} | x_{1:n}) = \prod_{t=1}^n p_\Theta(y_t | x_t)$$

- Para poder mejorar el uso de la notación, se omiten los insumos del modelo, pero también se tendrían que tener en cuenta
- Debido a que uno se centra en el modelo lineal gaussiano, si $g_p(x; \mu, \Sigma)$ denota la densidad normal multivariante con media μ y matriz de varianzas y covarianzas Σ , entonces se obtienen las siguientes equivalencias, con condiciones iniciales $g_p(x_0; \mu_0, \Sigma_0)$:

$$p_\Theta(x_t | x_{t-1}) = g_p(x_t; \Phi x_{t-1}, Q)$$

$$p_\Theta(y_t | x_t) = g_p(y_t; A_t x_t, R)$$

- En términos de densidades, el filtro de Kalman se puede entender como un esquema de actualización bayesiano, en el que, para determinar las densidades de las predicciones, se obtienen las siguientes igualdades:

$$\begin{aligned}
p_{\Theta}(\mathbf{x}_t | \mathbf{x}_{t-1}) &= \int_{\mathbb{R}^p} p_{\Theta}(\mathbf{x}_t, \mathbf{x}_{t-1} | \mathbf{y}_{1:t-1}) d\mathbf{x}_{t-1} = \\
&= \int_{\mathbb{R}^p} p_{\Theta}(\mathbf{x}_t | \mathbf{y}_{1:t-1}) p_{\Theta}(\mathbf{x}_{t-1} | \mathbf{y}_{1:t-1}) d\mathbf{x}_{t-1} = \\
&= \int_{\mathbb{R}^p} g_p(\mathbf{x}_t; \Phi \mathbf{x}_{t-1}, \mathbf{Q}) g_p(\mathbf{x}_{t-1}; \Phi \mathbf{x}_{t-1}^{t-1}, \mathbf{P}_{t-1}^{t-1}) d\mathbf{x}_{t-1} = \\
&= g_p(\mathbf{x}_t; \mathbf{x}_t^{t-1}, \mathbf{P}_t^{t-1})
\end{aligned}$$

- Estos valores se obtienen después de evaluar la integral usando el truco de completar el cuadrado
- Debido a que se busca un método iterativo, se ha introducido \mathbf{x}_{t-1} porque se presupone que se ha evaluado la densidad del filtro $p_{\Theta}(\mathbf{x}_{t-1} | \mathbf{y}_{1:t-1})$. Después de obtener el predictor, la densidad del filtro se obtiene de la siguiente manera, de la cual se deduce que es $g_p(\mathbf{x}_t; \mathbf{x}_t^t, \mathbf{P}_t^t)$:

$$\begin{aligned}
p_{\Theta}(\mathbf{x}_t | \mathbf{y}_{1:t}) &= p_{\Theta}(\mathbf{x}_t | \mathbf{y}_t, \mathbf{y}_{1:t-1}) \propto p_{\Theta}(\mathbf{y}_t | \mathbf{x}_t) p_{\Theta}(\mathbf{x}_t | \mathbf{y}_{1:t-1}) = \\
&= g_p(\mathbf{y}_t; \mathbf{A}_t \mathbf{x}_t, \mathbf{R}) g_p(\mathbf{x}_t; \Phi \mathbf{x}_{t-1}, \mathbf{Q}) = g_p(\mathbf{x}_t; \mathbf{x}_t^t, \mathbf{P}_t^t)
\end{aligned}$$

- Ahora, uno se enfoca en el problema de obtener estimadores para \mathbf{x}_t basados en la muestra entera de datos $\mathbf{y}_1, \dots, \mathbf{y}_n$, donde $t \leq n$, de modo que se quiere obtener \mathbf{x}_t^n
 - Estos estimadores se denominan suavizantes o *smoothers* porque un gráfico de la serie temporal $\{\mathbf{x}_t^n : t = 1, 2, \dots, n\}$ suele ser más suave que el de las predicciones $\{\mathbf{x}_t^{t-1} : t = 1, 2, \dots, n\}$ o el de los filtros $\{\mathbf{x}_t^t : t = 1, 2, \dots, n\}$
 - El suavizado implica que cada valor estimado sea una función del presente, del pasado y del futuro, mientras que un estimador para el filtro depende solo del presente y del pasado y el de las predicciones depende del pasado
- Para el modelo de estado-espacio especificado anteriormente, con condiciones iniciales \mathbf{x}_n^n y \mathbf{P}_n^n obtenidas a través del filtro de Kalman,

entonces las siguientes igualdades se mantienen para $t = n, n-1, \dots$, conocidas como el suavizante de Kalman:

$$x_{t-1}^n = x_{t-1}^{t-1} + J_{t-1}(x_t^n - x_t^{t-1}) \quad P_{t-1}^n = P_{t-1}^{t-1} + J_{t-1}(P_t^n - P_t^{t-1})J_{t-1}'$$

$$\text{where } J_{t-1} = P_{t-1}^{t-1} \Phi' (P_t^{t-1})^{-1}$$

- Para $t = 1, 2, \dots, n$ se define $y_{1:t-1} = \{y_1, y_2, \dots, y_{t-1}\}$ y $\eta_t = \{v_t, \dots, v_n, w_{t+1}, \dots, w_n\}$, con $y_{1:0}$ siendo un conjunto vacío, y se define $m_{t-1} = E(x_{t-1} | y_{1:t-1}, x_t - x_t^{t-1}, \eta_t)$
- Entonces, como $y_{1:t-1}$, $\{x_t - x_t^{t-1}\}$ y η_t son mutuamente independientes, y x_{t-1} y η_t son independientes, se puede obtener la siguiente igualdad a partir de la fórmula de la media condicional:

$$m_{t-1} = x_{t-1}^{t-1} + J_{t-1}(x_t - x_t^{t-1}) \quad \text{where}$$

$$J_{t-1} = Cov(x_{t-1}, x_t - x_t^{t-1})(P_t^{t-1})^{-1} = P_{t-1}^{t-1} \Phi' (P_{t-1}^{t-1})'$$

- Debido a que $y_{1:t-1}$, $\{x_t - x_t^{t-1}\}$ y η_t generan $y_{1:n} = \{y_1, y_2, \dots, y_n\}$, entonces se obtiene la primera igualdad vista anteriormente:

$$x_{t-1}^n = E(x_{t-1} | y_{1:n}) = E(m_{t-1} | y_{1:n}) =$$

$$= x_{t-1}^{t-1} + J_{t-1}(x_t^n - x_t^{t-1})$$

- Usando este resultado anterior y multiplicando cada lado por la su propia transpuesta, tomando esperanzas se puede obtener la siguiente igualdad:

$$x_{t-1} - x_{t-1}^n = x_{t-1} - x_{t-1}^{t-1} - J_{t-1}(x_t^n - x_t^{t-1})$$

$$\Rightarrow x_{t-1} - x_{t-1}^n = x_{t-1} - x_{t-1}^{t-1} - J_{t-1}(x_t^n - \Phi x_{t-1}^{t-1})$$

$$\Rightarrow (x_{t-1} - x_{t-1}^n) + J_{t-1} x_t^n = (x_{t-1} - x_{t-1}^{t-1}) + J_{t-1} \Phi x_{t-1}^{t-1}$$

$$\Rightarrow P_{t-1}^n + J_{t-1} E[x_t^n (x_t^n)'] J_{t-1}' = P_{t-1}^{t-1} + J_{t-1} \Phi E[x_{t-1}^{t-1} (x_{t-1}^{t-1})'] \Phi' J_{t-1}'$$

- Debido a que las esperanzas de la ecuación se pueden expresar en términos de P , se puede obtener la segunda ecuación ya vista:

$$E[x_t^n (x_t^n)'] = E[x_t x_t'] - P_t^n = \Phi E[x_{t-1} x_{t-1}'] \Phi' + Q - P_t^n$$

$$E \left[\mathbf{x}_{t-1}^{t-1} (\mathbf{x}_{t-1}^{t-1})' \right] = E[\mathbf{x}_{t-1} (\mathbf{x}_{t-1})'] - \mathbf{P}_{t-1}^{t-1}$$

$$\Rightarrow \mathbf{P}_{t-1}^n = \mathbf{P}_{t-1}^{t-1} + \mathbf{J}_{t-1} (\mathbf{P}_t^n - \mathbf{P}_t^{t-1}) \mathbf{J}'_{t-1}$$

- Para el modelo de estado-espacio anteriormente visto, con \mathbf{K}_t , \mathbf{J}_t y \mathbf{P}_n^n obtenidos a través del filtro y el suavizante de Kalman, y con condición inicial $\mathbf{P}_{n,n-1}^n = (\mathbf{I} - \mathbf{K}_n \mathbf{A}_n) \Phi \mathbf{P}_{n-1}^{n-1}$ para $t = n, n-1, n-2, \dots$, se puede obtener la siguiente igualdad, llamada *lag-one covariance smoother*:

$$\mathbf{P}_{t-1,t-2}^n = \mathbf{P}_{t-1}^{t-1} \mathbf{J}'_{t-2} + \mathbf{J}_{t-1} (\mathbf{P}_{t,t-1}^n - \Phi \mathbf{P}_{t-1}^{t-1}) \mathbf{J}'_{t-2}$$

- La estimación de los parámetros que especifica el modelo de estado-espacio es un poco complicada, pero normalmente se utiliza la estimación por máxima verosimilitud para poder obtener los parámetros
 - Se utiliza Θ para representar el vector de parámetros desconocidos en el vector de medias y la matriz de varianzas y covarianzas iniciales μ_0 y Σ_0 , la matriz de transición Φ y las matrices de varianzas y covarianzas de estado \mathbf{Q} y observación \mathbf{R} y las matrices de coeficientes de insumos \mathbf{Y} y Γ
 - Se utiliza la máxima verosimilitud bajo la suposición de que el estado inicial es normal $\mathbf{x}_0 \sim N_p(\mu_0, \Sigma_0)$ y los errores son normales, por lo que $\mathbf{w}_t \sim iid N_p(\mathbf{0}, \mathbf{Q})$ y $\mathbf{v}_t \sim iid N_p(\mathbf{0}, \mathbf{R})$
 - Por simplicidad, se suele asumir que $\{\mathbf{w}_t\}$ y $\{\mathbf{v}_t\}$ no están correlacionados
 - La verosimilitud se calcula utilizando las innovaciones $\epsilon_1, \epsilon_2, \dots, \epsilon_n$, definidas de la siguiente manera:

$$\epsilon_n = \mathbf{y}_t - \mathbf{A}_t \mathbf{x}_t^{t-1} - \Gamma \mathbf{u}_t$$

- La forma de las innovaciones de la verosimilitud de los datos $\mathbf{y}_{1:n}$, se obtiene usando un argumento similar al hecho para derivar la función de verosimilitud en el caso de los modelos ARMA, y procede señalando que las innovaciones son vectores aleatorios gaussianos independientes con media nula y con matriz de varianzas y covarianzas $\Sigma_t = \mathbf{A}_t \mathbf{P}_t^{t-1} \mathbf{A}'_t + \mathbf{R}$. Ignorando el término constante y multiplicando por -1 (para que se obtengan valores positivos y sea un problema de minimización), se obtiene la siguiente función:

$$-\ln L_Y(\Theta) = \frac{1}{2} \sum_{t=1}^n \ln |\Sigma_t(\Theta)| + \frac{1}{2} \sum_{t=1}^n \epsilon_t(\Theta)' \Sigma_t(\Theta)^{-1} \epsilon_t(\Theta)$$

- Esta función de los parámetros desconocidos es altamente no lineal y complicada, por lo que el procedimiento que se usa es uno basado en fijar valores iniciales x_0 (a través de una distribución normal multivariante), hacer recursiones y utilizar el método de Newton-Raphson
- Los pasos que se requieren para realizar el procedimiento de estimación con Newton-Raphson son los siguientes:
 - Se seleccionan variables para los parámetros, denotados por $\Theta^{(0)}$
 - Se ejecuta el filtro de Kalman usando los parámetros iniciales $\Theta^{(0)}$ para obtener un conjunto de innovaciones $\{\epsilon_t^{(0)} : t = 1, 2, \dots, n\}$ y de matrices de varianzas y covarianzas de los errores $\{\Sigma_t^{(0)} : t = 1, 2, \dots, n\}$
 - Se ejecuta una iteración del método de Newton-Rapshon con $-\ln L_Y(\Theta)$ como la función a minimizar para poder obtener un nuevo conjunto de parámetros $\Theta^{(1)}$
 - En la iteración j para $j = 1, 2, 3, \dots$, se repite el segundo paso usando $\Theta^{(j)}$ en lugar de $\Theta^{(j-1)}$ para obtener el nuevo conjunto de innovaciones $\{\epsilon_t^{(j)} : t = 1, 2, \dots, n\}$ y de matrices de varianzas y covarianzas de los errores $\{\Sigma_t^{(j)} : t = 1, 2, \dots, n\}$
 - Finalmente, se repite el tercer paso para obtener una nueva estimación $\Theta^{(j+1)}$. El algoritmo se detiene cuando las estimaciones o la verosimilitud se estabiliza, que se denota cuando $\Theta^{(j+1)}$ o $-\ln L_Y(\Theta^{(j+1)})$ difiere de $\Theta^{(j)}$ o $-\ln L_Y(\Theta^{(j)})$ por una pequeña cantidad predeterminada
- Además de este método, Shumway y Stoffer propusieron un método conceptualmente diferente que se basa en el algoritmo EM (el algoritmo *expectation-maximization*). Con tal de ser breve, uno puede ignorar los insumos u_t y se considera el modelo sin estos
 - La idea básica es que si se pudieran observar los estados $x_{0:n} = \{x_0, x_1, \dots, x_n\}$ además de las observaciones $y_{1:n} = \{y_1, \dots, y_n\}$, entonces se consideraría $\{x_{0:n}, y_{1:n}\}$ como los datos completos, con la siguiente densidad conjunta:

$$p_{\Theta}(x_{0:n}, y_{1:n}) = p_{\mu_0, \Sigma_0}(x_0) \prod_{t=1}^n p_{\Phi, Q}(x_t | x_{t-1}) \prod_{t=1}^n p_R(y_t | x_t)$$

- Debido a la suposición de normalidad e ignorando los términos constantes, la verosimilitud completa se puede escribir de la siguiente manera:

$$\begin{aligned} -2 \ln L_{X,Y}(\Theta) &= \ln |\Sigma_t(\Theta)| + (x_0 - \mu_0)' \Sigma_0^{-1} (x_0 - \mu_0) \\ &\quad + n \ln |Q| + \sum_{t=1}^n (x_t - \Phi x_{t-1})' Q^{-1} (x_t - \Phi x_{t-1}) \\ &\quad + n \ln |R| + \sum_{t=1}^n (y_t - A_t x_t)' R^{-1} (y_t - A_t x_t) \end{aligned}$$

- Por lo tanto, si se tuvieran los datos completos, se podrían utilizar las propiedades de la distribución normal multivariante para obtener los estimadores MLE de Θ
- No obstante, aunque no se pueden tener los datos completos, el algoritmo EM proporciona un método iterativo para encontrar los estimadores MLE de Θ basado en los datos incompletos $y_{1:n}$ al maximizar sucesivamente la expectación condicional en la verosimilitud de los datos completos
 - Para implementar este algoritmo, se escribe en cada iteración j para $j = 1, 2, 3, \dots$ la siguiente ecuación:

$$Q(\Theta | \Theta^{(j-1)}) = E[-2 \ln L_{XY}(\Theta) | y_{1:n}, \Theta^{(j-1)}]$$

- El cálculo de la ecuación anterior conforma el paso de la expectación o esperanza. Dado el valor actual $\Theta^{(j-1)}$, se puede usar el suavizante de Kalman para obtener las esperanzas condicionales como suavizados, y permite obtener las siguientes ecuaciones:

$$\begin{aligned} Q(\Theta | \Theta^{(j-1)}) &= \ln |\Sigma_t(\Theta)| + \text{tr} [\Sigma_0^{-1} (P_0^n + (x_0^n - \mu_0)(x_0^n - \mu_0)')] \\ &\quad + n \ln |Q| \\ &\quad + \text{tr} [Q^{-1} (S_{11} - S_{10}\Phi' - \Phi S_{10}' + \Phi S_{00}\Phi')] \\ &\quad + n \ln |R| \\ &\quad + \text{tr} \left[R^{-1} \sum_{t=1}^n [(y_t - A_t x_t^n)(y_t - A_t x_t^n)' + A_t P_t^n A_t'] \right] \end{aligned}$$

where

$$S_{00} = \sum_{t=1}^n [x_{t-1}^n (x_{t-1}^n)' + P_{t-1}^n]$$

$$S_{10} = \sum_{t=1}^n [x_t^n (x_{t-1}^n)' + P_{t,t-1}^n]$$

$$S_{11} = \sum_{t=1}^n [x_t^n (x_t^n)' + P_t^n]$$

- Los suavizantes se calculan bajo el valor corriente de los parámetros $\Theta^{(j-1)}$, aunque no se ha mostrado este hecho de manera explícita. Además, al obtener $Q(\cdot | \cdot)$, se hace uso de $E(x_s x_t' | y_{1:n}) = x_s^n (x_t^n)' + P_{s,t}^n$ (no se reemplaza x_t por x_t^n en la verosimilitud y ya)
- Minimizar la expresión $Q(\Theta | \Theta^{(j-1)})$ con respecto a los parámetros, en la iteración j , corresponde al paso de maximización, y es análogo al enfoque de la regresión multivariante, que permite obtener las siguientes estimaciones:

$$\Phi^{(j)} = S_{10} S_{00}^{-1}$$

$$Q^{(j)} = n^{-1} (S_{11} - S_{10} S_{00}^{-1} S_{10}')$$

$$R^{(j)} = n^{-1} \sum_{t=1}^n [(y_t - A_t x_t^n) (y_t - A_t x_t^n)' + A_t P_t^n A_t']$$

- Las actualizaciones que se obtienen para el vector de medias inicial y la matriz de varianzas y covarianzas inicial son las siguientes, obtenidas al minimizar $Q(\Theta | \Theta^{(j-1)})$:

$$\mu_0^{(j)} = x_0^n \quad \Sigma_0^{(j)} = P_0^n$$

- El procedimiento general se puede interpretar como una alternación entre las recursiones del filtro y del suavizado de Kalman y los estimadores de máxima verosimilitud de la distribución normal multivariante. Este proceso se puede resumir en los siguientes pasos:

- Se inicializa el algoritmo escogiendo valores iniciales para los parámetros en $\{\mu_0, \Sigma_0, \Phi, Q, R\}$ (denotado por Θ_0) y se calcula la verosimilitud para los datos incompletos – $\ln L_Y(\Theta_0)$
- Para la iteración j de $j = 1, 2, 3, \dots$, se realiza el paso E: usando los parámetros $\Theta^{(j-1)}$, se usa el filtrado y el suavizado de Kalman

para obtener los valores suavizados x_t^n , P_t^n y $P_{t,t-1}^n$ para $t = 1, \dots, n$ y se calcula S_{00} , S_{10} y S_{11}

- Para la iteración j de $j = 1, 2, 3, \dots$, se realiza el paso M: se actualizan los parámetros $\{\mu_0, \Sigma_0, \Phi, Q, R\}$ usando las fórmulas resultantes de la minimización, obteniendo así $\Theta^{(j)}$
- Para la iteración j de $j = 1, 2, 3, \dots$, se calcula la verosimilitud de los datos incompletos – $\ln L_Y(\Theta^{(j)})$
- Se repiten los pasos del segundo al cuarto hasta que el algoritmo converja
- La distribución asintótica de los estimadores de los parámetros $\widehat{\Theta}_n$ se suele estudiar en términos muy generales, y la consistencia y la normalidad asintótica de los estimadores se establece bajo condiciones muy generales
 - Una condición esencial para esto es la estabilidad del filtro, la cual asegura que para una gran t , las innovaciones ϵ_t son básicamente copias unas de las otras con una matriz de varianzas y covarianzas estable Σ que no depende de t y que, asintóticamente, las innovaciones contienen toda la información sobre los parámetros desconocidos
 - Aunque no es necesario, por simplicidad, se debería asumir a partir de ahora que $A_t = A$ para toda t . No obstante, desviaciones de este supuesto se encuentran en varias referencias
 - Bajo condiciones generales, siendo $\widehat{\Theta}_n$ el estimador de Θ_0 obtenido por la maximización de las innovaciones de la verosimilitud, $L_Y(\Theta)$ (dada por la verosimilitud de los datos incompletos), entonces se cumple la siguiente convergencia:

$$\sqrt{n}(\widehat{\Theta}_n - \Theta_0) \xrightarrow{D} N(\mathbf{0}, \mathfrak{T}(\Theta_0)^{-1})$$

$$\text{where } \mathfrak{T}(\Theta) = \lim_{n \rightarrow \infty} n^{-1} E \left[-\frac{\partial^2 \ln L_Y(\Theta)}{\partial \Theta \partial \Theta'} \right]$$

- Para un procedimiento de Newton, la matriz Hessiana en el momento de convergencia se puede usar como un estimador de $n\mathfrak{T}(\Theta_0)$ para obtener estimadores de los errores estándar
- En el caso del algoritmo EM, no se calculan derivadas, pero se puede incluir una evaluación numérica de la matriz Hessiana en el momento de la convergencia para obtener los errores estándar

Los modelos de estado-espacio: datos perdidos y modelos estructurales

- Una característica atractiva del marco de los modelos de estado-espacio es la capacidad de tratar las series temporales que se han observado irregularmente en el tiempo, habiendo la posibilidad de representar modelos ARFIMA
 - Suponiendo que en un momento t , se define la partición del vector de observaciones $q \times 1$ en dos partes: $\mathbf{y}_t^{(1)}$, el vector de valores observados de tamaño $q_{1t} \times 1$, y $\mathbf{y}_t^{(2)}$, el vector de valores no observados de tamaño $q_{2t} \times 1$, en donde $q_{1t} + q_{2t} = q$
 - De este modo, la ecuación de observaciones partida será la siguiente, donde $\mathbf{A}_t^{(1)}$ y $\mathbf{A}_t^{(2)}$ son, respectivamente, las matrices de observación de tamaño $q_{1t} \times p$ y $q_{2t} \times p$:

$$\begin{pmatrix} \mathbf{y}_t^{(1)} \\ \mathbf{y}_t^{(2)} \end{pmatrix} = \begin{bmatrix} \mathbf{A}_t^{(1)} \\ \mathbf{A}_t^{(2)} \end{bmatrix} \mathbf{x}_t + \begin{pmatrix} \mathbf{v}_t^{(1)} \\ \mathbf{v}_t^{(2)} \end{pmatrix}$$

- Además, la matriz de varianzas y covarianzas del vector de los errores de medición se puede expresar de la siguiente manera:

$$\mathbf{R}_t = \begin{pmatrix} \mathbf{v}_t^{(1)} \\ \mathbf{v}_t^{(2)} \end{pmatrix} = \begin{bmatrix} \mathbf{R}_{11t} & \mathbf{R}_{12t} \\ \mathbf{R}_{21t} & \mathbf{R}_{22t} \end{bmatrix}$$

- En el caso en el que hay datos perdidos y $\mathbf{y}_t^{(2)}$ no se observa, se puede modificar la ecuación de observación en el modelo de estado-espacio especificado anteriormente para obtener el siguiente modelo:

$$\mathbf{x}_t = \Phi \mathbf{x}_{t-1} + \mathbf{w}_t \quad \mathbf{y}_t^{(1)} = \mathbf{A}_t^{(1)} \mathbf{x}_t + \mathbf{v}_t^{(1)}$$

- En este caso, la ecuación de observación es una ecuación vectorial $q_{1t} \times 1$ en el momento t
- Debido a que, por el corolario anteriormente visto, las ecuaciones del filtro de Kalman se mantienen (con las substituciones de notación adecuadas) para parámetros que varían en el tiempo. Si no hay observaciones en el momento t , entonces se fija la matriz de ganancias de Kalman $\mathbf{K}_t = \mathbf{0}$, de modo que $\mathbf{x}_t^t = \mathbf{x}_t^{t-1}$ y $\mathbf{P}_t^t = \mathbf{P}_t^{t-1}$
- En vez de utilizar dimensiones observacionales que varían, es más fácil computacionalmente modificar el modelo fijando a $\mathbf{0}$ ciertos parámetros y retener una ecuación de observaciones q -dimensional

- En particular, el corolario sobre el filtro de Kalman aplica para el caso con datos perdidos si, en la actualización t , se sustituyen los siguientes vectores por \mathbf{y}_t , \mathbf{A}_t y \mathbf{R} , en donde \mathbf{I}_{22t} es una matriz identidad $q_{2t} \times q_{2t}$

$$\mathbf{y}_{(t)} = \begin{pmatrix} \mathbf{y}_t^{(1)} \\ \mathbf{0} \end{pmatrix} \quad \mathbf{A}_{(t)} = \begin{bmatrix} \mathbf{A}_t^{(1)} \\ \mathbf{0} \end{bmatrix} \quad \mathbf{R}_{(t)} = \begin{bmatrix} \mathbf{R}_{11t} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_{22t} \end{bmatrix}$$

- En este caso, los valores para el proceso de innovaciones tendrán la siguiente forma:

$$\boldsymbol{\epsilon}_{(t)} = \begin{pmatrix} \boldsymbol{\epsilon}_t^{(1)} \\ \mathbf{0} \end{pmatrix} \quad \boldsymbol{\Sigma}_{(t)} = \begin{bmatrix} \mathbf{A}_t^{(1)} \mathbf{P}_t^{t-1} (\mathbf{A}_t^{(1)})' + \mathbf{R}_{11t} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_{22t} \end{bmatrix}$$

- Esto último hace que la forma en términos de las innovaciones de la verosimilitud sea correcta para este caso, por lo que las sustituciones hacen que las estimaciones por máxima verosimilitud a través de la función en términos de las innovaciones puedan realizarse como en el caso en el que los datos están completos
- Una vez los valores filtrados de las observaciones perdidas se han obtenido, los valores suavizados se pueden procesar utilizando el suavizante de Kalman y el *lag-one covariance smoother* con los valores obtenidos de los valores perdidos filtrados con los datos
 - En el caso de datos perdidos, los estimadores de estado se denotan de la siguiente manera, y las covarianzas del *lag-one smoother* se denotan por $\mathbf{P}_{t,t-1}^{(n)}$:

$$\mathbf{x}_t^{(s)} = E(\mathbf{x}_t | y_1^{(s)}, y_2^{(s)}, \dots, y_s^{(s)})$$

$$\mathbf{P}_t^{(s)} = E[(\mathbf{x}_t - \mathbf{x}_t^{(s)})(\mathbf{x}_t - \mathbf{x}_t^{(s)})']$$

- Los estimadores máximos verosímiles en el procedimiento EM requiere más modificaciones. Considerando $\mathbf{y}_{1:n}^{(1)} = \{y_1^{(1)}, \dots, y_n^{(1)}\}$ como los datos incompletos, y $\{\mathbf{x}_{0:n}, \mathbf{y}_{1:n}\}$ como los datos completos, la verosimilitud de los datos completos será la misma que antes, pero para el paso de la esperanza, en la iteración j , se tiene que hacer el siguiente cálculo:

$$Q(\Theta | \Theta^{(j-1)}) = E[-2 \ln(L_{X,Y}) | \mathbf{y}_{1:n}^{(1)}, \Theta^{(j-1)}] =$$

$$= E_{\Theta^{(j-1)}} [\ln |\boldsymbol{\Sigma}_0| + \text{tr}(\boldsymbol{\Sigma}_0^{-1}(\mathbf{x}_0 - \boldsymbol{\mu}_0)(\mathbf{x}_0 - \boldsymbol{\mu}_0)') | \mathbf{y}_{1:n}^{(1)}] +$$

$$E_{\Theta^{(j-1)}} \left[n \ln |\mathbf{Q}| + \sum_{t=1}^n [tr(\mathbf{Q}^{-1}(\mathbf{x}_t - \Phi \mathbf{x}_{t-1})(\mathbf{x}_t - \Phi \mathbf{x}_{t-1})')] \mid \mathbf{y}_{1:n}^{(1)} \right] +$$

$$+ E_{\Theta^{(j-1)}} \left[n \ln |\mathbf{R}| + \sum_{t=1}^n [tr(\mathbf{R}^{-1}(\mathbf{y}_t - \mathbf{Ax}_t)(\mathbf{y}_t - \mathbf{Ax}_t)')] \mid \mathbf{y}_{1:n}^{(1)} \right]$$

- Los primeros dos términos en la ecuación son como los primeros dos términos de $Q(\Theta | \Theta^{(j-1)})$ cuando se usa estimación áximo verosímil con los suavizantes $\mathbf{x}_t^n, \mathbf{P}_t^n$ y $\mathbf{P}_{t,t-1}^n$ reemplazados por sus contrapartes con datos perdidos $\mathbf{x}_t^{(n)}, \mathbf{P}_t^{(n)}$ y $\mathbf{P}_{t,t-1}^{(n)}$
- El tercer término se debe de evaluar más específicamente $E_{\Theta^{(j-1)}} \left[\mathbf{y}_t^{(2)} \mid \mathbf{y}_{1:n}^{(1)} \right]$ y $E_{\Theta^{(j-1)}} \left[\mathbf{y}_t^{(2)} (\mathbf{y}_t^{(2)})' \mid \mathbf{y}_{1:n}^{(1)} \right]$, lo cual permite obtener el siguiente resultado, donde $\mathbf{R}_{*,ikt}$ para $i, k = 1, 2, \mathbf{x}_t^{(n)}$ y $\mathbf{P}_t^{(n)}$ indican los valores actuales especificados por $\Theta^{(j-1)}$:

$$\begin{aligned} & E_{\Theta^{(j-1)}} \left[(\mathbf{y}_t - \mathbf{Ax}_t)(\mathbf{y}_t - \mathbf{Ax}_t)' \mid \mathbf{y}_{1:n}^{(1)} \right] = \\ & = \left(\begin{matrix} \mathbf{y}_t^{(1)} - \mathbf{A}_t^{(1)} \mathbf{x}_t^{(n)} \\ \mathbf{R}_{*,21t} \mathbf{R}_{*,11t}^{-1} (\mathbf{y}_t^{(1)} - \mathbf{A}_t^{(1)} \mathbf{x}_t^{(n)}) \end{matrix} \right) \left(\begin{matrix} \mathbf{y}_t^{(1)} - \mathbf{A}_t^{(1)} \mathbf{x}_t^{(n)} \\ \mathbf{R}_{*,21t} \mathbf{R}_{*,11t}^{-1} (\mathbf{y}_t^{(1)} - \mathbf{A}_t^{(1)} \mathbf{x}_t^{(n)}) \end{matrix} \right)' \\ & + \left(\begin{matrix} \mathbf{A}_t^{(1)} \\ \mathbf{R}_{*,21t} \mathbf{R}_{*,11t}^{-1} \mathbf{A}_t^{(1)} \end{matrix} \right) \mathbf{P}_t^{(n)} \left(\begin{matrix} \mathbf{A}_t^{(1)} \\ \mathbf{R}_{*,21t} \mathbf{R}_{*,11t}^{-1} \mathbf{A}_t^{(1)} \end{matrix} \right)' \\ & + \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{R}_{*,22t} - \mathbf{R}_{*,21t} \mathbf{R}_{*,11t}^{-1} \mathbf{R}_{*,12t} \end{pmatrix} \end{aligned}$$

- En el caso en el que no hay correlación entre los errores de los componentes observados y los componentes no observados, entonces la fórmula anterior se simplifica al siguiente caso:

$$\begin{aligned} & E_{\Theta^{(j-1)}} \left[(\mathbf{y}_t - \mathbf{Ax}_t)(\mathbf{y}_t - \mathbf{Ax}_t)' \mid \mathbf{y}_{1:n}^{(1)} \right] = \\ & = (\mathbf{y}_{(t)} - \mathbf{A}_{(t)} \mathbf{x}_t^{(n)}) (\mathbf{y}_{(t)} - \mathbf{A}_{(t)} \mathbf{x}_t^{(n)})' + \mathbf{A}_{(t)} \mathbf{P}_t^{(n)} \mathbf{A}_{(t)}' + \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{R}_{*,22t} \end{pmatrix} \end{aligned}$$

- En este caso simplificado, el paso M con datos perdidos es similar al caso anteriormente visto sin datos perdidos. Primero, se obtienen los siguientes parámetros:

$$S_{(11)} = \sum_{t=1}^n \left[\mathbf{x}_t^{(n)} (\mathbf{x}_t^{(n)})' + \mathbf{P}_t^{(n)} \right]$$

$$S_{(10)} = \sum_{t=1}^n \left[\mathbf{x}_t^{(n)} \left(\mathbf{x}_{t-1}^{(n)} \right)' + \mathbf{P}_{t,t-1}^{(n)} \right]$$

$$S_{(00)} = \sum_{t=1}^n \left[\mathbf{x}_{t-1}^{(n)} \left(\mathbf{x}_{t-1}^{(n)} \right)' + \mathbf{P}_{t-1}^{(n)} \right]$$

- En la iteración j , por tanto, el paso de maximización (paso M) es el siguiente, donde \mathbf{D}_t es una matriz de permutación que reordena las variables en el momento t en su orden original:

$$\Phi^{(j)} = S_{(10)} S_{(00)}^{-1}$$

$$\mathbf{Q}^{(j)} = n^{-1} (S_{(10)} S_{(00)}^{-1} S'_{(10)})$$

$$\mathbf{R}^{(j)} = n^{-1} \sum_{t=1}^n \mathbf{D}_t \left(E_{\Theta^{(j-1)}} \left[(\mathbf{y}_t - \mathbf{A}\mathbf{x}_t)(\mathbf{y}_t - \mathbf{A}\mathbf{x}_t)' \middle| \mathbf{y}_{1:n}^{(1)} \right] \right) \mathbf{D}_t'$$

- En la iteración j , solo se actualizaría R_{11t} , mientras que R_{22t} se quedaría en su valor estimado para la iteración $j-1$ (se recuerda que R_{*22t} es la matriz estimada dado $\Theta^{(j-1)}$). Si no se puede asumir que $R_{12} = \mathbf{0}$, entonces $\mathbf{R}^{(j)}$ se tiene que cambiar usando $E_{\Theta^{(j-1)}} \left[(\mathbf{y}_t - \mathbf{A}\mathbf{x}_t)(\mathbf{y}_t - \mathbf{A}\mathbf{x}_t)' \middle| \mathbf{y}_{1:n}^{(1)} \right]$ en el caso general, pero $\Phi^{(j)}$ y $\mathbf{Q}^{(j)}$ se quedan igual
- Igual que antes, las estimaciones de los parámetros para el estado inicial se actualizan de la siguiente manera:

$$\mu_0^{(j)} = \mathbf{x}_0^{(n)} \quad \Sigma_0^{(j)} = \mathbf{P}_0^{(n)}$$

- Los modelos estructurales son modelos de componentes en los que cada componente explica un tipo específico de comportamiento, y estos modelos se pueden ajustar fácilmente al marco de los modelos de estado-espacio
 - Los modelos normalmente son alguna versión de la descomposición clásica de las series temporales en componentes de tendencia, estacionalidad e irregularidad
 - Consecuentemente, cada componente tiene una interpretación directa como la naturaleza de la variación en los datos
 - Asumiendo una descomposición aditiva clásica para las series temporales, es posible suponer una forma funcional para los componentes de tendencia y de estacionalidad

$$y_t = \mu_t + S_t + v_t$$

$$\mu_t = \phi\mu_{t-1} + w_{t1} \quad S_t + S_{t-1} + S_{t-2} + S_{t-3} = w_{t2}$$

- La ecuación para el componente de tendencia indica que el crecimiento de esta es exponencial con $\phi > 1$, mientras que la ecuación del componente estacional es un modelo AR no estacionario (con retrasos con coeficientes unitarios) cuya suma esperada es cero
- Para expresar el modelo en forma de estado-espacio, se considera un vector de estado $x_t = (\mu_t, S_t, S_{t-1}, S_{t-2})'$, de manera que la ecuación de observación se puede expresar de la siguiente manera:

$$y_t = (1, 1, 0, 0) \begin{pmatrix} \mu_t \\ S_t \\ S_{t-1} \\ S_{t-2} \end{pmatrix} + v_t$$

- La ecuación de estado y la matriz Q y R para los errores w_t , en este caso, se puede expresar de la siguiente manera:

$$\begin{pmatrix} \mu_t \\ S_t \\ S_{t-1} \\ S_{t-2} \end{pmatrix} = \begin{bmatrix} \phi & 0 & 0 & 0 \\ 0 & -1 & -1 & -1 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{pmatrix} \mu_{t-1} \\ S_{t-1} \\ S_{t-2} \\ S_{t-3} \end{pmatrix} + \begin{pmatrix} w_{t1} \\ w_{t2} \\ 0 \\ 0 \end{pmatrix}$$

$$R = r_{11} \quad Q = \begin{bmatrix} q_{11} & 0 & 0 & 0 \\ 0 & q_{22} & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

- El modelo se reduce la forma de estado espacio con $p = 4$ y $q = 1$
- Una manera alternativa de realizar una descomposición estructural de una serie temporal es a través de considerar un componente estacional y una tendencia local

$$y_t = \mu_t + S_t + v_t$$

$$\begin{cases} \mu_t = \mu_{t-1} + \beta_{t-1} + w_{t1} \\ \beta_t = \beta_{t-1} + w_{t2} \end{cases} \quad S_t + S_{t-1} + S_{t-2} + S_{t-3} = w_{t3}$$

- Para expresar el modelo en forma de estado-espacio, se considera un vector de estado $x_t = (\mu_t, \beta_t, S_t, S_{t-1}, S_{t-2})'$, de manera que la ecuación de observación se puede expresar de la siguiente manera:

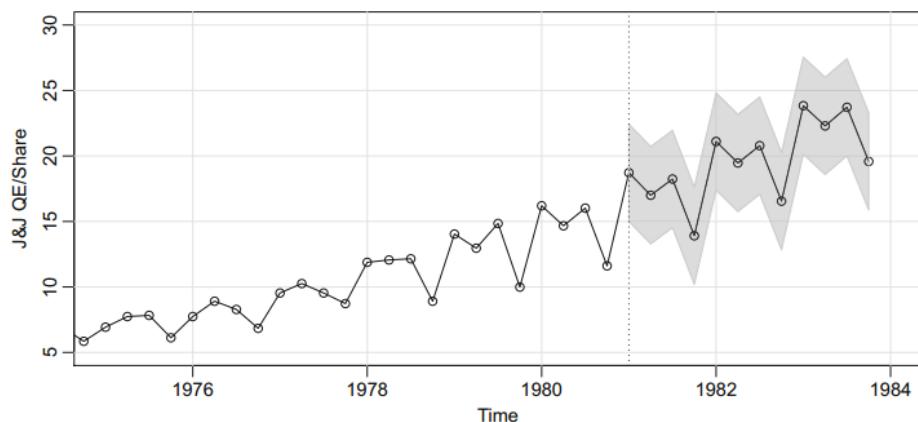
$$y_t = (1, 0, 1, 0, 0) \begin{pmatrix} \mu_t \\ \beta_t \\ S_t \\ S_{t-1} \\ S_{t-2} \end{pmatrix} + v_t$$

- La ecuación de estado y la matriz \mathbf{Q} y \mathbf{R} para los errores \mathbf{w}_t , en este caso, se puede expresar de la siguiente manera:

$$\begin{pmatrix} \mu_t \\ \beta_t \\ S_t \\ S_{t-1} \\ S_{t-2} \end{pmatrix} = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & -1 & -1 & -1 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix} \begin{pmatrix} \mu_{t-1} \\ \beta_{t-1} \\ S_{t-1} \\ S_{t-2} \\ S_{t-3} \end{pmatrix} + \begin{pmatrix} w_{t1} \\ w_{t2} \\ w_{t3} \\ 0 \\ 0 \end{pmatrix}$$

$$\mathbf{R} = r_{11} \quad \mathbf{Q} = \begin{bmatrix} q_{11} & 0 & 0 & 0 & 0 \\ 0 & q_{22} & 0 & 0 & 0 \\ 0 & 0 & q_{33} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

- Por supuesto, se puede dar muchas especificaciones alternativas para los componentes, pero lo importante es modelar los componentes de la serie acorde al comportamiento observado (tendencia global o local, patrón estacional constante o cambiante, etc.)
- Los parámetros a estimar son r_{11} , la varianza del ruido en las ecuaciones de medida, q_{11} y q_{12} , las varianzas del modelo correspondiendo a los componentes de tendencia y de estacionalidad y el parámetro de transición que modela la tasa de crecimiento ϕ



- Este modelo se puede estimar a través del método de Newton-Raphson

Los modelos de estado-espacio: errores correlacionados y bootstrap

- Hay veces que es ventajoso escribir el modelo de estado espacio de una manera diferente, con tal de poder trabajar con el marco de estos modelos en contextos especiales
 - Un modelo de estado espacio se puede escribir de la siguiente manera:

$$\boldsymbol{x}_{t+1} = \Phi \boldsymbol{x}_t + \boldsymbol{\Upsilon} \boldsymbol{u}_{t+1} + \boldsymbol{\Theta} \boldsymbol{w}_t$$

- En la ecuación de estado, $\boldsymbol{x}_0 \sim N_p(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$ es el punto inicial, Φ es una matriz $p \times p$, $\boldsymbol{\Upsilon}$ es una matriz $p \times r$, $\boldsymbol{\Theta}$ es una matriz $p \times m$ y $\boldsymbol{w}_t \sim iid N_m(0, Q)$. En cambio, en la ecuación de observación, \boldsymbol{A}_t es una matriz $q \times p$, $\boldsymbol{\Gamma}$ es una matriz $q \times r$ y $\boldsymbol{v}_t \sim iid N_q(0, R)$
- En este modelo, mientras que w_t y v_t aún son series de ruido blanco (independientes de \boldsymbol{x}_0), pero se permite que los ruidos del estado y de la observación estén correlacionados. Esto se modela a través de la covarianza, donde \boldsymbol{S} es una matriz $m \times q$ y δ_s^t es la delta de Kronecker

$$Cov(w_s, v_t) = \boldsymbol{S} \delta_s^t \quad \text{where } \delta_s^t = \begin{cases} 1 & \text{if } s = t \\ 0 & \text{if } s \neq t \end{cases}$$

- La mayor diferencia entre esta forma del modelo y la que se ha visto anteriormente es que el proceso del ruido comienza en $t = 0$ con tal de mejorar la notación relacionada con la concurrencia de la covarianza entre w_t y v_t . Además, la inclusión de la matriz $\boldsymbol{\Theta}$ permite que se evite usar un proceso de un estado singular
- Para obtener las innovaciones $\boldsymbol{\varepsilon}_t = \boldsymbol{y}_t - \boldsymbol{A}_t \boldsymbol{x}_t^{t-1} - \boldsymbol{\Gamma} \boldsymbol{u}_t$ y la varianza de las innovaciones $\boldsymbol{\Sigma}_t = \boldsymbol{A}_t \boldsymbol{P}_t^{t-1} \boldsymbol{A}_t' + \boldsymbol{R}$, en este caso, se necesitan las predicciones a un paso adelante del estado
 - Las estimaciones filtradas también serán de interés, y serán necesarias para el suavizado, cuyas ecuaciones anteriores se siguen aplicando en este caso
 - La siguiente propiedad genera el predictor \boldsymbol{x}_{t+1}^t del predictor pasado \boldsymbol{x}_t^{t-1} cuando los términos de ruido están correlacionados y exhiben una actualización de filtro

- Para el modelo de estado-espacio especificado, con condiciones iniciales \mathbf{x}_1^0 y \mathbf{P}_1^0 para $t = 1, 2, \dots, n$ y donde $\boldsymbol{\varepsilon}_t = \mathbf{y}_t - \mathbf{A}_t \mathbf{x}_t^{t-1} - \boldsymbol{\Gamma} \mathbf{u}_t$, se pueden obtener las siguientes ecuaciones:

$$\mathbf{x}_{t+1}^t = \Phi \mathbf{x}_t^{t-1} + \Upsilon \mathbf{u}_{t+1} + K_t \boldsymbol{\varepsilon}_t$$

$$\mathbf{P}_{t+1}^t = \Phi \mathbf{P}_t^{t-1} \Phi' + \Theta \mathbf{Q} \Theta' - K_t \Sigma_t K_t'$$

- En este caso, la matriz de la ganancia de Kalman se da por la siguiente expresión:

$$K_t = (\Phi \mathbf{P}_t^{t-1} \mathbf{A}'_t + \Theta \mathbf{S}) (\mathbf{A}_t \mathbf{P}_t^{t-1} \mathbf{A}'_t + R)^{-1}$$

- Los valores filtrados se dan por las siguientes ecuaciones:

$$\mathbf{x}_t^t = \mathbf{x}_t^{t-1} + \mathbf{P}_t^{t-1} \mathbf{A}'_t (\mathbf{A}_t \mathbf{P}_t^{t-1} \mathbf{A}'_t + R)^{-1} \boldsymbol{\varepsilon}_t$$

$$\mathbf{P}_t^t = \mathbf{P}_t^{t-1} - \mathbf{P}_t^{t-1} \mathbf{A}'_{t+1} (\mathbf{A}_t \mathbf{P}_t^{t-1} \mathbf{A}'_t + R)^{-1} \mathbf{A}_t \mathbf{P}_t^{t-1}$$

- La derivación de estos resultados es similar a la derivación del filtro de Kalman visto anteriormente, aunque la matriz de ganancia K_t difiere. Los valores filtrados son simbólicamente idénticos a los vistos anteriormente
- Para inicializar el filtro, se utilizan las siguientes ecuaciones:

$$\mathbf{x}_1^0 = E(\mathbf{x}_1) = \Phi \boldsymbol{\mu}_0 + \Upsilon \mathbf{u}_1$$

$$\mathbf{P}_1^0 = Var(\mathbf{x}_1) = \Phi \Sigma_0 \Phi' + \Theta \mathbf{Q} \Theta'$$

- El modelo anterior se puede usar para modelos ARMAX y para modelos de regresión multivariante con errores correlacionados, siendo posible combinar los modelos. Para el ARMAX, los insumos entran en la ecuación de estado y para la regresión entran en la de observación
 - Considerando un modelo ARMAX k -dimensional, las observaciones \mathbf{y}_t son un proceso vectorial k -dimensional, Φ y Θ son matrices $k \times k$, Υ es una matriz $k \times r$, \mathbf{u}_t es el insumo de r -dimensional y $\boldsymbol{\nu}_t$ es un proceso de ruido blanco vectorial $k \times 1$

$$\mathbf{y}_t = \Upsilon \mathbf{u}_t + \sum_{j=1}^p \Phi_j \mathbf{y}_{t-j} + \sum_{k=1}^q \Theta_k \boldsymbol{\nu}_{t-k} + \boldsymbol{\nu}_t$$

- Este modelo es igual al visto anteriormente, pero se escriben las observaciones como \mathbf{y}_t

- Definiendo las matrices \mathbf{F} de tamaño $kp \times kp$, \mathbf{G} de tamaño $kp \times k$ y \mathbf{H} de tamaño $kp \times r$, el modelo de estado espacio se puede definir de la siguiente manera, donde $\mathbf{A} = (\mathbf{I}, \mathbf{0}, \dots, \mathbf{0})$ es una matriz $k \times pk$ e \mathbf{I} es la matriz identidad de tamaño $k \times k$:

$$\mathbf{F} = \begin{bmatrix} \Phi_1 & \mathbf{I} & \mathbf{0} & \cdots & \mathbf{0} \\ \Phi_1 & \mathbf{0} & \mathbf{I} & \cdots & \mathbf{0} \\ \cdots & \cdots & \cdots & \cdots & \mathbf{0} \\ \Phi_{p-1} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{I} \\ \Phi_p & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} \end{bmatrix} \quad \mathbf{G} = \begin{bmatrix} \Theta_1 + \Phi_1 \\ \cdots \\ \Theta_q + \Phi_q \\ \Phi_{q+1} \\ \cdots \\ \Phi_p \end{bmatrix} \quad \mathbf{H} = \begin{bmatrix} \Upsilon \\ \mathbf{0} \\ \cdots \\ \mathbf{0} \end{bmatrix}$$

$$\mathbf{x}_{t+1} = \mathbf{F}\mathbf{x}_t + \mathbf{H}\mathbf{u}_{t+1} + \mathbf{G}\mathbf{v}_t$$

$$\mathbf{y}_t = \mathbf{A}\mathbf{x}_t + \mathbf{v}_t$$

- Si $p < q$, entonces $\Phi_{p+1} = \cdots = \Phi_q = \mathbf{0}$, por lo que $q = p$. El proceso de estado de este modelo es kp -dimensional, mientras que el de las observaciones es k -dimensional
- Considerando un modelo univariante $ARMAX(1,1)$ donde $\alpha_t = \Upsilon u_t$, se pueden usar los resultados vistos anteriormente para poder obtener el modelo de estado-espacio de la siguiente manera:

$$y_t = \alpha_t + \phi y_{t-1} + \theta v_{t-1} + v_t$$

$$\Rightarrow x_{t+1} = \phi x_t + \alpha_{t+1} + (\theta + \phi)v_t$$

$$\Rightarrow y_t = x_t + v_t$$

- En este caso, $w_t = v_t$ en la ecuación de estado, por lo que $Cov(w_t, v_t) = Var(v_t) = R$ y $Cov(w_t, v_s) = 0$ cuando $s \neq t$ y eso hace que se aplique el filtro de Kalman cuando hay errores correlacionados
- Para pasar del modelo de estado-espacio al ARMAX, se pueden hacer las siguientes transformaciones:

$$\begin{aligned} y_t &= x_t + v_t = \phi x_{t-1} + \alpha_t + (\theta + \phi)v_{t-1} + v_t = \\ &= \alpha_t + \phi(x_{t-1} + v_{t-1}) + v_t + \theta v_{t-1} = \\ &= \alpha_t + \phi y_{t-1} + v_t + \theta v_{t-1} \end{aligned}$$

- Se pueden usar el modelo en forma de estado-espacio con el filtro de Kalman con errores correlacionados para conseguir la estimación de máxima verosimilitud para los modelos ARMAX vistos anteriormente

Los modelos de estado-espacio: modelos lineales dinámicos y volatilidad estocástica

- Los modelos de volatilidad estocástica son una alternativa a los modelos tipo GARCH, y estos se pueden expresar en forma de modelo de estado-espacio
 - En estos modelos, r_t denota los rendimientos de un activo financiero. La mayoría de modelos para datos de rendimientos usados en la práctica tienen una forma multiplicativa como $r_t = \sigma_t \varepsilon_t$
 - En esta ecuación, ε_t es una secuencia iid y el proceso de volatilidad σ_t es un proceso estocástico no negativo tal que ε_t es independiente de σ_s para toda $s \leq t$
 - Normalmente se asume que ε_t con media nula y con varianza unitaria
- En los modelos de volatilidad estocástica, la volatilidad es una transformación no lineal de un proceso autorregresivo lineal escondido, en donde el proceso escondido $x_t = \log(\sigma_t^2)$ sigue la siguiente autorregresión de primer orden:

$$x_t = \phi x_{t-1} + w_t \quad r_t = \beta \exp\left(\frac{x_t}{2}\right) \varepsilon_t$$

- En este caso, $w_t \sim iid N(0, \sigma_w^2)$ y ε_t es un ruido iid que tiene momentos finitos. Los procesos de error w_t y ε_t se asumen como mutuamente independientes y $|\phi| < 1$
- Como w_t es normalmente distribuidos, x_t también es normal, y todos los momentos de ε_t existen, de modo que los momentos de r_t también existen
- Asumiendo que $x_0 \sim N(0, \sigma_x^2/(1 - \phi^2))$ (la distribución estacionaria), la curtosis de r_t se da por $\kappa_4(r_t) = \kappa_4(\varepsilon_t) \exp(\sigma_x^2)$, donde $\sigma_x^2 = \sigma_w^2/(1 - \phi^2)$ es la varianza estacionaria de x_t
- La función de autocorrelación de $\{r_t^{2m}; t = 1, 2, \dots\}$ para cualquier número entero m se da por la siguiente fórmula:

$$\text{corr}(r_{t+h}^{2m}, r_t^{2m}) = \frac{\exp(m^2 \sigma_x^2 \phi^h) - 1}{\kappa_4(\varepsilon_t) \exp(m^2 \sigma_x^2) - 1}$$

- La tasa de caída o *decay rate* de la función de autocorrelación es más rápida que la exponencial en retrasos de un tiempo pequeño y se estabiliza en ϕ para retrasos grandes
- A veces es más fácil trabajar con una forma lineal del modelo en donde se define $y_t = \log(r_t^2)$, $x_t = \log(\sigma_t^2)$ y $v_t = \log(\varepsilon_t^2)$. Por lo tanto, se puede escribir y_t de la siguiente manera:

$$y_t = \alpha + x_t + v_t$$

- Esta ecuación normalmente proviene de considerar un proceso de los rendimientos multiplicado por una constante positiva (interpretada como la desviación estándar marginal), como $r_t = c\sigma_t\varepsilon_t$, de modo que $\log(r_t^2) = \log(c^2) + \log(\sigma_t^2) + \log(\varepsilon_t^2)$
- Una constante normalmente se necesita en la ecuación de estado o en la ecuación de observaciones (no en ambas), por lo que se puede escribir la ecuación de estado de la siguiente manera:

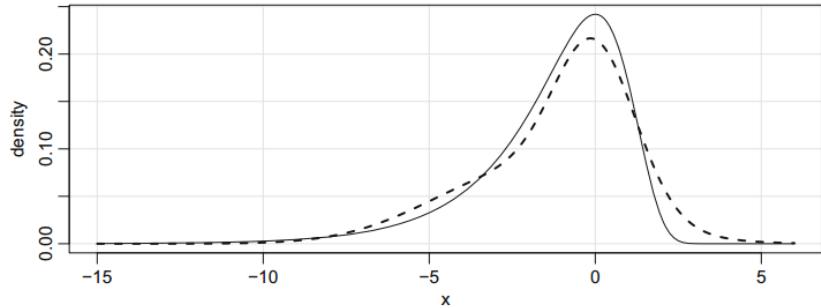
$$x_t = \phi_0 + \phi_1 x_{t-1} + w_t$$

- En esta ecuación, w_t es ruido blanco gaussiano con varianza σ_w^2 . La constante ϕ_0 normalmente se conoce como el efecto apalancamiento o *leverage effect*
- Juntando las ecuaciones para y_t y para x_t , se obtiene el modelo de volatilidad estocástica de Taylor
- Si ε_t^2 tuviera una distribución log-normal, las ecuaciones formarían un modelo de estado-espacio gaussiano y se podrían utilizar resultados del modelo de estado-espacio desarrollados anteriormente para ajustar el modelo
 - No obstante, esta suposición no funciona en el modelo. Si, en vez de esto, se mantiene la suposición de normalidad del ARCH sobre $\varepsilon_t \sim iid N(0,1)$, en cuyo caso, $v_t = \log(\varepsilon_t^2)$ se distribuye como el logaritmo de una variable aleatoria distribuida como una chi cuadrada con un grado de libertad, cuya densidad viene dada por la siguiente función:

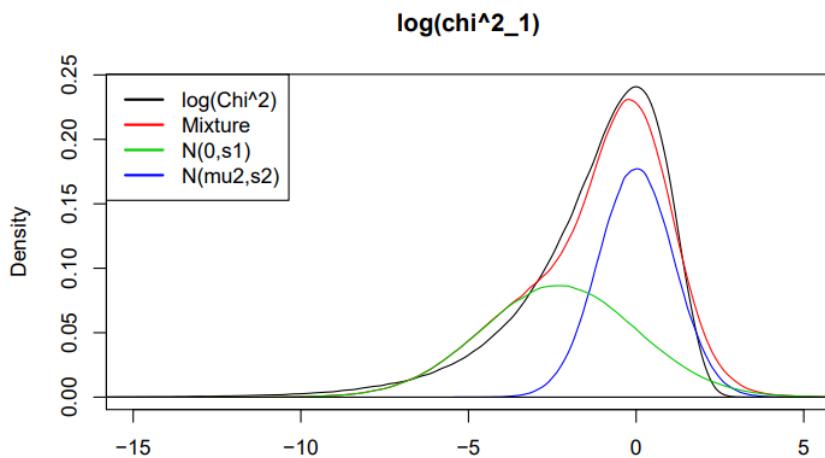
$$f(v) = \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{1}{2}(e^v - v)\right] \quad for \quad -\infty < v < \infty$$

- La media de esta distribución es $-(\gamma + \log(2))$, donde γ es la constante de Euler, y la varianza de la distribución es $\pi^2/2$. Esta

densidad está muy sesgada pero no es flexible porque no hay parámetros libres para estimar



- Muchos enfoques para el ajuste de los modelos de volatilidad estocástica se han examinado, pero estos incluyen muchas suposiciones sobre el proceso de ruido en la ecuación de observación
- Un modelo propuesto por Kim, Shepard y Chib es modelar el logaritmo de la variable aleatoria distribuida como una chi cuadrada por una mezcla de siete distribuciones normales para aproximar los primeros cuatro momentos de la distribución del error observacional
 - De esta manera, la mezcla es fija y no se añaden parámetros adicionales para el modelo se añaden al utilizar esta técnica
- La suposición de que ε_t es gaussiano no es realista para la mayoría de aplicaciones, por lo que, en un esfuerzo de mantener las cosas simples, pero de manera general (las dinámicas del error observacional dependen en parámetros ajustados), el método para ajustar el modelo de volatilidad estocástica es el siguiente:



- Se mantiene la ecuación de estado gaussiana, pero se utiliza un ruido blanco η_t cuya distribución es una mezcla de dos normales y una se centra en cero

$$y_t = \alpha + x_t + \eta_t$$

$$x_t = \phi_0 + \phi_1 x_{t-1} + w_t$$

$$\eta_t = I_t z_{t0} + (1 - I_t) z_{t1}$$

- En este caso, I_t es un proceso de Bernoulli iid con $P(I_t = 0) = \pi_0$ y $P(I_t = 1) = \pi_1$ (tal que $\pi_0 + \pi_1 = 1$), $z_{t0} \sim iid N(0, \sigma_0^2)$ y $z_{t1} \sim iid N(\mu_1, \sigma_1^2)$
- En este modelo, los parámetros a estimar son $\phi_0, \phi_1, Q, \alpha, \sigma_0, \mu_2$ y σ_1
- La ventaja de este modelo es que es fácil de ajustar debido a que utiliza normalidad. Y se pueden utilizar las siguientes ecuaciones de filtro:

$$x_{t+1}^t = \phi_0 + \phi_1 x_t^{t-1} + \sum_{j=0}^1 \pi_{tj} K_{tj} \varepsilon_{tj}$$

$$P_{t+1}^t = \phi_1^2 P_t^{t-1} + \sigma_w^2 - \sum_{j=0}^1 \pi_{tj} K_{tj}^2 \Sigma_{tj}$$