

Departamento de Estadística
Facultad de Ciencias
Sede Bogotá



UNIVERSIDAD
NACIONAL
DE COLOMBIA

Notas de Clase Series de Tiempo

Sergio Alejandro Calderón Villanueva

Departamento de Estadística
Facultad de Ciencias
Universidad Nacional de Colombia
Bogotá D.C, Colombia
2023

Contenido

1	Introducción al Análisis de Series de Tiempo	3
1.1	Aspectos del análisis de series de tiempo	3
1.2	Un ejemplo de regresión	9
2	Procesos Estocásticos y Procesos Lineales	15
2.1	Procesos estocásticos Estacionarios	16
2.1.1	Procesos Lineales	28
2.2	Análisis Descriptivo de una Series de Tiempo	38
2.2.1	Importación de datos en R y Python	39
2.2.2	Gráfico de una Serie de Tiempo	39
2.2.3	Análisis de Tendencias	40
2.2.4	Suavizamiento de Series de Tiempo	44
2.2.5	Transformación Box-Cox	48
2.2.6	Diagramas de dispersión para la variable y sus retardos y el ACF muestral.	51
2.2.7	Detección de Ciclos y Estacionalidades	55
3	Procesos ARMA	63
3.1	Expansión de funciones Racionales	68
3.1.1	Relaciones de Recurrencia	69
3.2	El modelo ARMA(1,1)	71
3.3	Características de los Procesos ARMA(p,q) Generales	74
3.4	Forma de Computar la Función de Autocovarianza de un pro- ceso ARMA	76

3.5	Predicción de Procesos Estacionarios	80
3.5.1	Las ecuaciones de predicción en el dominio del tiempo	84
3.6	Función de Autocorrelación Parcial	90
4	Estimación de la media y funciones de autocovarianza auto-correlación parcial de un Proceso Estocástico Estacionario	95
4.1	Estimación de la Media	95
4.2	Estimación de la función de autocovarianza y de autocorrelación	99
5	Estimación de Procesos ARMA	109
5.1	Estimación Vía Máxima Verosimilitud	109
5.2	Función de Verosimilitud de un Proceso ARMA(p,q)	110
5.3	Propiedades Asintóticas de los estimadores de máxima verosimilitud	112
5.4	Identificación de los ordenes p,q y análisis de residuales	115
5.4.1	Criterios de Información	115
5.4.2	Comprobación de los supuestos(Diagnóstico basados en los residuales)	117
6	Modelo ARIMA para Series de Tiempo No Estacionarias	121
6.1	Modelo Autoregresivos Integrados y de Promedio Móviles(ARIMA)	122
6.2	Modelamiento de Series no Estacionarias Vía Modelos ARIMA	128
6.3	Pronósticos con Modelos ARIMA	130
6.4	Series de Tiempo Estacionales y Modelos SARIMA	132
7	Metodologías Alternativas para Series de Tiempo No Estacionarias	139
7.1	Descomposición por Filtros de Promedios Móviles	140
7.2	Suavizamiento Exponencial	142
7.3	Rolling y Evaluación de los pronósticos	144
7.3.1	Medidas de Precisión.	146
7.4	Modelos Estructurales de Series de Tiempo	146
7.5	Predicción de las Componentes del Modelo Estructural	151
7.6	Preliminares del Aprendizaje Automático	154
7.6.1	Teoría Estadística de la Decisión	155
7.6.2	Evaluación de Algoritmos	158
7.6.3	Enfoques estándar para la determinación de algoritmos	164
7.6.4	Entrenamiento, Validación y Prueba	167

7.6.5	Métodos de validación alternativos	171
7.6.6	Procesamiento de datos en la práctica	177
7.7	Árboles, Redes Neuronales y Aprendizaje Profundo	179
7.7.1	Métodos basados en Árboles	180
7.7.2	Árboles de decisión	180
7.7.3	Poda del Árbol	184
7.7.4	Aprendizaje Automático	185
7.7.5	Redes Neuronales Multicapa	188
7.8	Redes Neuronales para Múltiple-Respuesta de respuesta Vectorial	203
7.9	Combinación de Pronósticos	207
8	Datos Atípicos, Datos Faltantes y Análisis Espectral de Series de Tiempo.	211
8.1	Análisis de Intervención	211
8.2	Datos Atípicos	213
8.3	Estimación o Predicción de Observaciones Faltantes	218
8.3.1	Estimación de datos faltantes usando el Filtro de Kalman	218
8.3.2	Estimación usando análisis de intervención	219
8.4	Análisis Espectral de Series de Tiempo	219
9	Modelo de Heterocedasticidad Condicional	223
9.1	Serie de Retornos	223
10	Series de tiempo Multivariadas	229
10.1	Preliminares	229
10.2	Procesos Estacionarios	233
10.2.1	Estimación del vector de medias y la FACV matricial	239
10.2.2	Prueba Portmanteau para Correlación Cruzada Cero .	243
10.2.3	Pronóstico	243
10.3	Regresión de Series de Tiempo	245
10.3.1	Regresión Lineal simple	245
10.3.2	Regresión Lineal Múltiple	246
10.3.3	Relaciones Espurias	247
10.3.4	Otros Regresores útiles	247
10.3.5	Selección de predictores	250
10.3.6	Pronósticos con modelos de Regresión	250
10.4	Modelos de Regresión Dinámica	252
10.4.1	Pronóstico	253

10.4.2	Predictores retardados	253
10.5	Árboles de Regresión y Bosques Aleatorios	254
10.5.1	Crecimiento	254
10.5.2	Poda del Árbol	256
10.5.3	Bosques Aleatorios	257
10.5.4	Bagging	257
10.6	Aspectos a tener en cuenta en un modelo de regresión	259
11 Procesos VARMA		261
11.1	Supuestos Básicos y propiedades de los Procesos VAR	261
11.1.1	Procesos Estables VAR(p)	261
11.1.2	Cómputo de la la función de autocovarianza y de autocorrelación matricial de un proceso VAR estable	268
11.1.3	Análisis Estructural con Modelos VAR	276
11.2	Estimación de Procesos Autorregresivo Vectoriales	282
11.3	Estimadores de Máxima Verosimilitud	286
11.4	Pronóstico con Procesos Estimados	290
11.5	Selección y Diagnóstico del Modelo	291
11.5.1	Criterios de Selección del Modelo	291
11.6	Prueba de Causalidad de Granger	296
11.6.1	Prueba de Causalidad de Granger del tipo Wald	296
12 Procesos Integrados y Modelos de Corrección de Errores Vectoriales		299
12.1	Regresión Espuria	299
12.2	Combinaciones Lineales de un Proceso Vectorial	300
12.3	Co-integración	301
12.3.1	Sobre-diferenciación	301
12.4	Modelos de Corrección de Errores Vectoriales(VEC)	303
12.5	Prueba de Co-integración	305
12.6	Estimación de Modelo de Corrección de Errores	307
12.7	Pronóstico de Procesos Integrados y Co-integrados	308
13 Modelo Factorial Dinámico		309
13.1	DFM para Series Estacionarias	309
13.2	Factores Dinámicos y Modelos VARMA	314
13.3	Ajuste de un DFM estacionario a los datos	315
13.4	Componentes Principales	319
13.4.1	PCA Estático	319

13.4.2	Propiedades de las PC	321
13.5	Componentes Principales Dinámicas(PCA Dinámico(DPCA))	322
13.5.1	DPCs de un lado o cara	323
13.5.2	Selección del Modelo y Pronóstico	324
14	Agrupamiento o Clustering de Series de Tiempo	327
14.1	Distancias y Disimilaridades	328
14.1.1	Distancias entre Series de Tiempo	328
14.1.2	Disimilaridades entre Series de Tiempo Univariadas .	330
14.2	Agrupamiento o Aglomeración Jerárquico	331
14.2.1	Criterio para Definir Distancias entre Grupos	331
14.2.2	Selección del número de Grupos	332

Introducción

Estas notas de clase se han elaborado para desarrollar principalmente metodología y teoría estadística en series de tiempo univariadas. Estas notas presentan tanto modelos lineales como procesos no-lineales. También habrá una sesión donde introduciremos algunos métodos de aprendizaje automático para realizar pronósticos de series de tiempo. También hablaremos acerca de un modelo de series de tiempo para el caso donde el objeto de estudio es una función, es decir modelo de series de tiempo para datos funcionales. Finalmente se incluyen códigos en software como R, Python.

1

Introducción al Análisis de Series de Tiempo

En este capítulo daremos una introducción acerca de los que llamaremos el análisis de series de tiempo, iniciando con la idea intuitiva de que es una serie de tiempo, para luego presentar diversos ejemplos de series de tiempo resaltando las características que ellas presentan, al igual que un ejemplo del análisis de regresión de series temporales.

1.1 Aspectos del análisis de series de tiempo

Inicialmente daremos una definición intuitiva de lo que es una serie de tiempo. Consideremos a X como la variable aleatoria de interés. Esta variable puede ser por ejemplo la precipitación en un región específica; o el producto interno bruto de un país; o por ejemplo el precio de acción específica. Si las mediciones de esta variable de interés son hechas en un periodo de tiempo, digamos desde el tiempo $t = 1$ hasta el tiempo $t = T$, entonces obtenemos las realizaciones $\{x_1, \dots, x_T\}$ de la variable de interés. A éstas realizaciones la llamaremos una *serie de tiempo*. Por ejemplo, en la figura 1.1 tenemos la gráfica a través del tiempo de la variable de interés X , que es la precipitación medida mensualmente para Colombia, la cual es obtenida a través de un promedio ponderado por departamento.

Otro ejemplo de una serie tiempo es la del índice COLCAP, el cual es computado usando un promedio ponderado del valor de las 20 acciones más líquidas que se cotizan en la bolsa de valores de Colombia. Una gráfica de ésta serie de tiempo es mostrada en la figura 1.2. Note que esta serie muestra una tendencia, la cual es en general creciente aunque no monotonamente, no

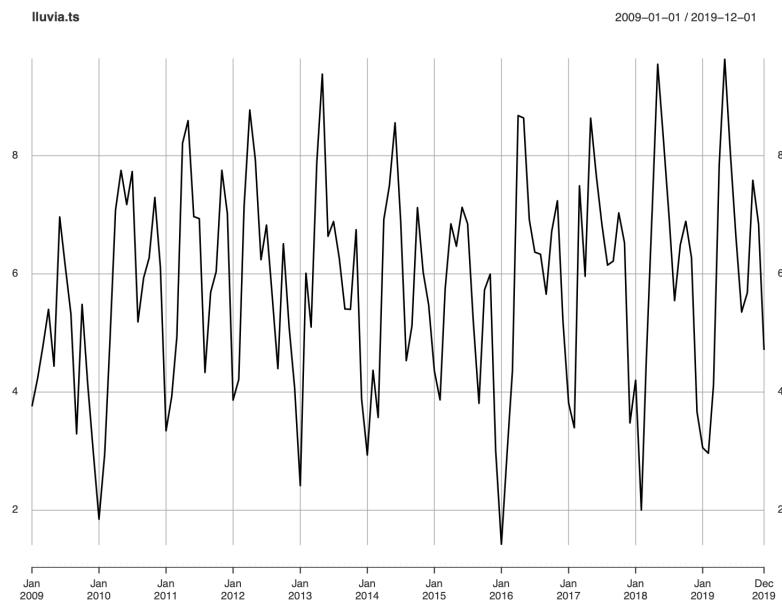


Figura 1.1: Lluvias mensuales ponderadas para Colombia desde Enero de 2009 hasta Diciembre de 2019

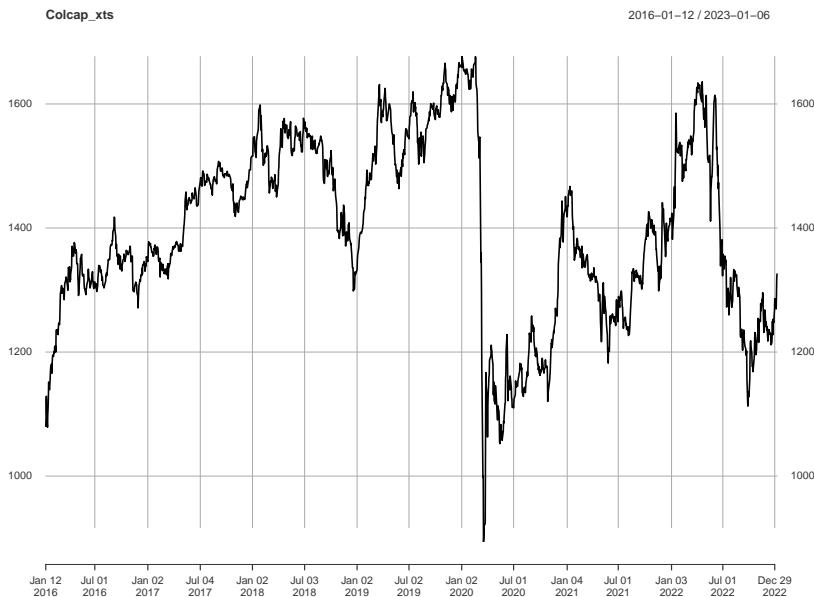


Figura 1.2: Índice COLCAP desde el 12 de Enero de 2016 hasta el 6 de Enero de 2023

ea clara si es una tendencia determinística o estocástica. No parece tener un ciclo estacional, ni tampoco heterocedasticidad marginal.

Veamos otras dos series de tiempo, la tasa de desempleo en Colombia mensual y la demanda de energía diaria en Colombia.

Entonces podemos asumir que teóricamente tenemos el siguiente escenario desde el punto de vista probabilístico: Dada una variable de interés X que es medida desde el tiempo $t = 1$ hasta el tiempo $t = T$, entenderemos a X_t como la variable aleatoria medida en el tiempo t . Si por ejemplo tenemos que nuestra variable de interés es la precipitación mensual para Colombia, entonces

X_1 : Precipitación en Colombia en el mes 1 $\rightarrow x_1$: realización de la V.A. X_1

X_2 : Precipitación en Colombia en el mes 2 $\rightarrow x_2$: realización de la V.A. X_2
 \vdots

X_T : Precipitación en Colombia en el mes $T \rightarrow x_T$: realización de la V.A. X_T
Este esquema nos lleva pensar entonces en algunos aspectos:

- 1.) Las variables aleatorias X_1, \dots, X_T pueden presentar alguna estructura

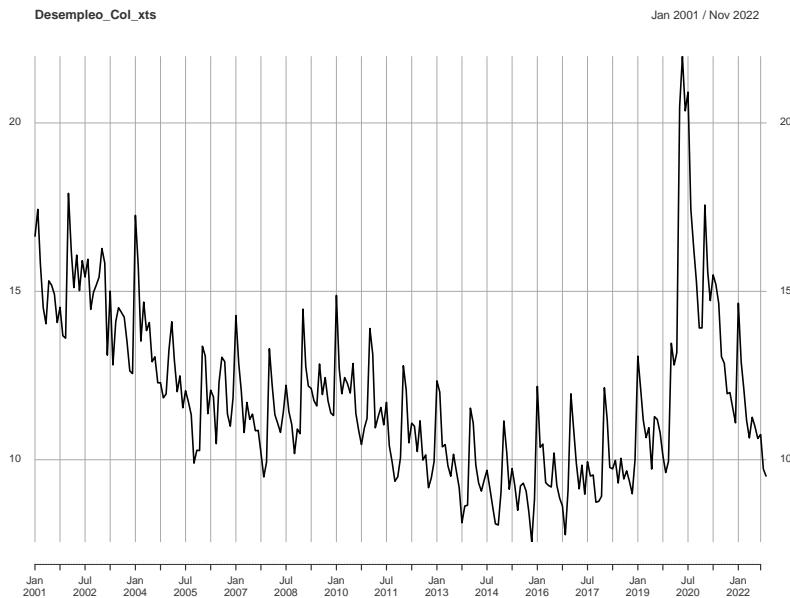


Figura 1.3: Tasa de Desempleo Mensual Colombia

de autocorrelación y por lo tanto el supuesto de independencia que se venía asumiendo en muestras aleatorias ya no se mantiene.

- 2.) Asumir que las unidades experimentales que generaron la serie de tiempo $\{x_t\}$ son diferentes ya no es necesariamente cierto.
- 3.) Debido algunos comportamientos o características que pueden presentar las series de tiempo, el supuesto de igual distribución para las variables aleatorias puede no mantenerse.

Por otro lado, al chequear la gráfica de una serie de tiempo, se podría observar algunas de las siguientes características o patrones en ellas, por ejemplo:

Tendencia: es un comportamiento determinístico o aleatorio el cual hace que las observaciones no fluctúen alrededor de un valor medio específico fijo. La serie COLCAP presenta tendencia, mientras que la serie de lluvias no parece presentar una tendencia.

Círculo estacional: el valor medio depende del mes o del día considerado, en un periodo fijo o regular .

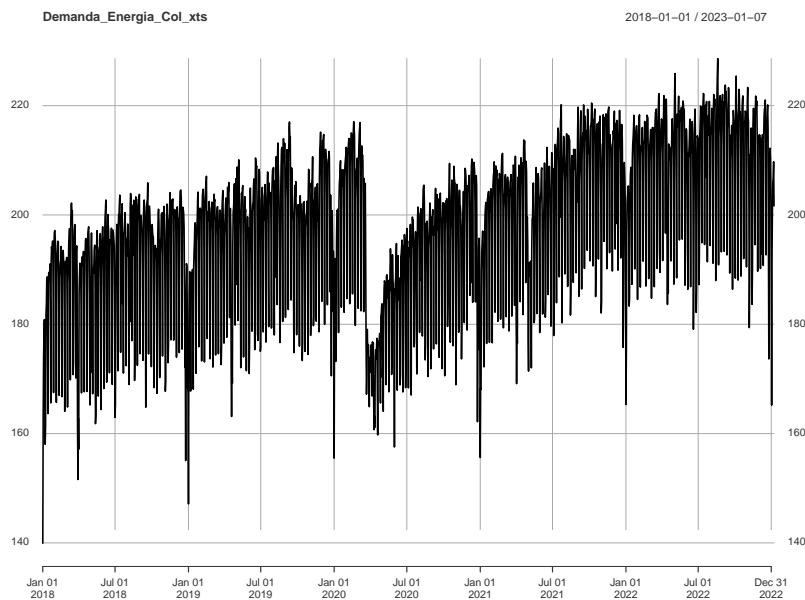


Figura 1.4: Demanda Diaria de Energía Colombia

Cíclico: El valor medio puede depender no necesariamente del mismo mes o día, en un periodo no necesariamente fijo, el cual es ,mas largo que el ciclo estacional.

Heterocedasticidad: La variabilidad marginal no es constante a lo largo del tiempo, es decir, el el recorrido de la variable se amplía o disminuye con el tiempo. Sin embargo podría también presentarse heterocedasticidad condicional.

Nota 1.1. Una razonable suposición para el modelamiento de nuestra series de tiempo $\{Y_t\}$ es que las componentes o patrones se pueden escribir de forma aditiva como sigue:

$$Y_t = T_t + S_t + X_t,$$

donde T_t : componente de tendencia, S_t : componente estacional y X_t : componente de error aleatorio.

La serie de pasajeros es una de las series mas famosas y se encuentra como ejemplo en todos los textos de series de tiempo.

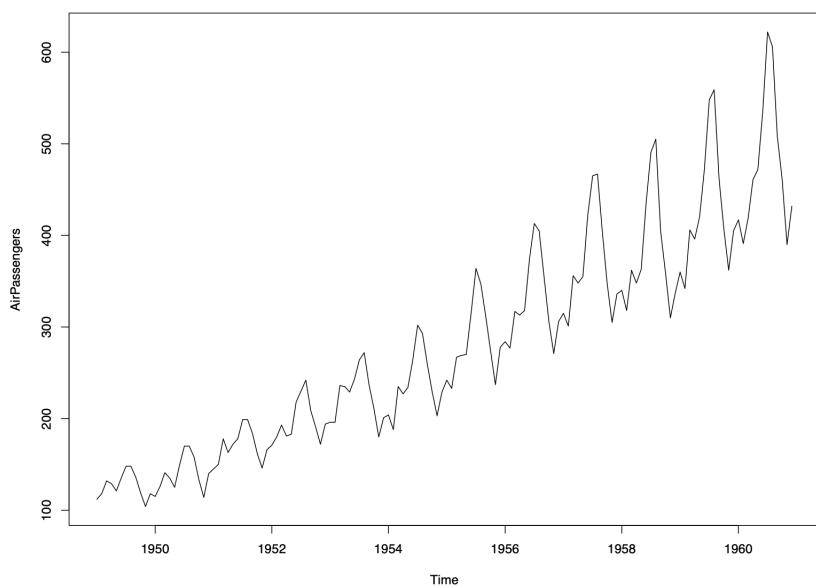


Figura 1.5: Serie Mensual del número de pasajeros desde Enero de 1949 hasta Diciembre de 1960

Las anteriores características y otras mas podrían presentarse en una serie de tiempo. La idea en un análisis de series de tiempo desde un enfoque estadístico consiste entonces en establecer uno o varios modelos probabilísticos que representen los datos, es decir, que refleje las características que presentan los datos. Note por ejemplo que la serie mensual del número de pasajeros de una aerolínea que se muestra en la figura 1.5 muestra algunas de las anteriores características mencionadas. Entonces, el modelo probabilístico mencionado se conocerá como un modelo de series de tiempo o para la serie de tiempo. Así tenemos en seguida nuestra primera definición:

Definición 1.2. *Un modelo de Series de Tiempo para los datos observados $\{x_t\}$ es una especificación de las distribuciones conjuntas(o posiblemente sólo las medias y covarianzas) de una sucesión de variables aleatorias $\{X_t\}$ (proceso estocástico) de las cuales $\{x_t\}$ se postula sea un realización truncada o parcial.*

Note que la definición anterior es bastante general, lo cual sugiere que el modelamiento se pueda abordar desde un enfoque paramétrico o no paramétrico. Adicionalmente en la definición, podemos observar que se sugiere abordar el tema de procesos estocásticos, por lo tanto, en el siguiente capítulo abordaremos éste tema.

Finalmente, debemos mencionar que los objetivos del análisis de series de tiempo desde un enfoque Estadístico que desarrollaremos a lo largo de éste escrito son los siguientes:

- (a.) Establecer un modelo probabilístico que represente los datos.(se debe estudiar las características que presentan las series de tiempo.)
- (b.) Estimar los parámetros del modelo propuesto.(asumiendo que nos estamos basando en un enfoque paramétrico)
- (c.) Comprobar la bondad del ajuste del modelo a los datos.(verificación de supuestos)
- (d.) Usar el modelo ajustado ya sea para entender el comportamiento de los datos o para hacer pronósticos.

1.2 Un ejemplo de regresión

En finanzas es usual la búsqueda de relaciones entre variables, un ejemplo de esto es el modelo de mercado que relaciona el exceso de retorno de una

acción individual (exceso de retorno de la acción de la GM) a un índice del mercado (S&P 500). Este tipo de modelo es conocido como el CAPM, y se utiliza ampliamente en el sector financiero para fijar el precio de valores riesgosos y generar los rendimientos esperados de los activos dado el riesgo de esos activos y el costo de capital. Otro ejemplo consiste en la estructura de tasas de interés, la cual busca la evolución en el tiempo de la relación entre las tasas de interés con diferentes vencimientos (tasas de interés semanal con vencimientos a 1 año 3 años). Este modelo es usado para identificar el estado actual de la economía. La estructura temporal de las tasas de interés refleja las expectativas de los participantes del mercado sobre cambios futuros en las tasas de interés y su evaluación de las condiciones de política monetaria. Ver libro [36] Capítulo 2 y <https://www.investopedia.com> para las definiciones de los términos. Para estos ejemplos, un modelo razonable sería

$$y_t = \alpha + \beta x_t + e_t \quad (1.1)$$

donde y_t y x_t representan las variables medidas en el tiempo t , y e_t representa el término de error. Si e_t es IID, entonces el método de mínimos cuadrados ordinarios puede ser usado para estimar α , β y σ_e^2 . Vamos a introducir un modelo lineal para para identificar si existe relación entre las tasas de interés semanales en Estados Unidos:

- r_{1t} : la tasa de interés del Tesoro a vencimiento constante a 1 año.
- r_{3t} : la tasa de interés del Tesoro a vencimiento constante a 3 años.

Las observaciones van desde el 5 de enero de 1962, hasta el 10 de abril de 2009, para un total de 2467 observaciones Ver ejemplo mas detallado en [36]. Veamos las gráficas en el tiempo de las variables originales 1.6 y en cambios 1.7, es decir, $c_{1t} = r_{1t} - r_{1,t-1}$ y $c_{3t} = r_{3t} - r_{3,t-1}$:

Vamos a explorar las relaciones de las variables mediante el gráfico de dispersión de las tasas, y de los cambios de las tasas 1.8 :

Note que tanto las tasas, como los cambios de las tasas parecen tener una relación bastante fuerte 1.8. Veamos el ajuste de los modelos en el Markdown de R IntroSeriesUnivariadas.Rmd, allí podemos ver que el ajuste del modelo de regresión simple entre las tasas $r_{3t} = \alpha + \beta r_{1t} + e_t$ nos muestra que existe relación entre las dos variables, pero dicho ajuste no es adecuado porque al validar los supuestos, ellos no se cumplen.

Se hace un nuevo ajuste, aunque ahora con las series de cambios de las tasas $c_{3t} = \beta c_{1t} + e_t$ debido al comportamiento no estable a través del tiempo tanto

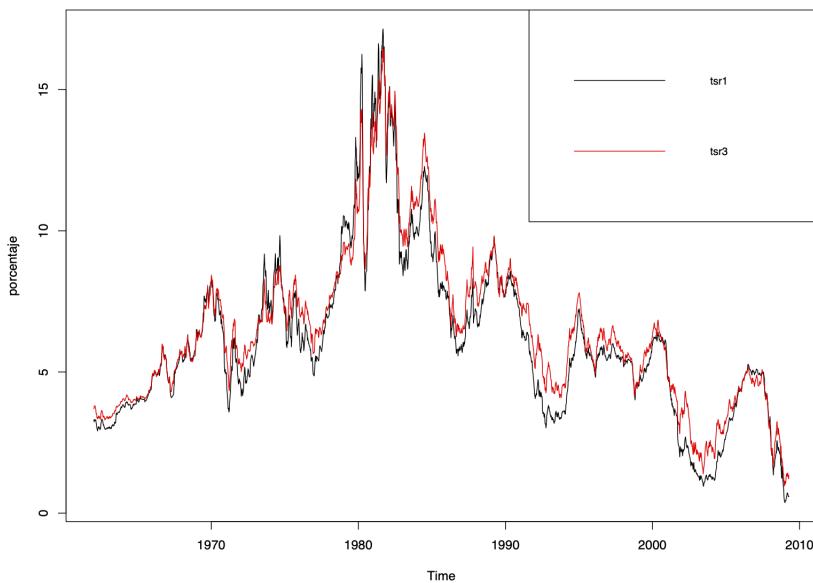


Figura 1.6: Gráficas de las variables originales

de las tasas como de los los residuales. Ver el mismo archivo de Rmarkdown. Finalmente, un modelo de series de tiempo parece mas razonable con el objeto de tener en cuenta la estructura de autocorrelación presente en los residuales. Por supuesto, un enfoque equivalente sería considerar un modelo lineal usual, pero el ajuste debería hacerse vía mínimos cuadrados generalizados ($\hat{\beta} = (X'W^{-1}X)^{-1}X'W^{-1}y$, y $Var(\hat{\beta}) = (X'W^{-1}X)^{-1}$) donde W es la matriz de covarianza del error. Se puede usar el comando *gls* de R en el paquete *nlme*. Los resultados están también en R Markdown.

Tarea 1.3. Considera la carpeta *Base de Datos de repositorio de Github* de la rama de *Series Univariadas*. Use el archivo *DesempleoEmpleo.xlsx*, tome la variable *Tasa de empleo* y haga la gráfica de la misma contra el tiempo. Tiene sentido analizar esta gráfica desde el enfoque de series de tiempo, explique. Realice lo mismo con los archivos *Heatrow.xlsx* y *Centralpark.xlsx* relacionadas con variables del clima.

Recuerde que debe seleccionar dos tipos de series de tiempo para el proyecto. Estas series deben ser de frecuencias de observación distintas. Una de las dos debe tener tendencia. Al menos una de las dos debe tener una componente estacional.

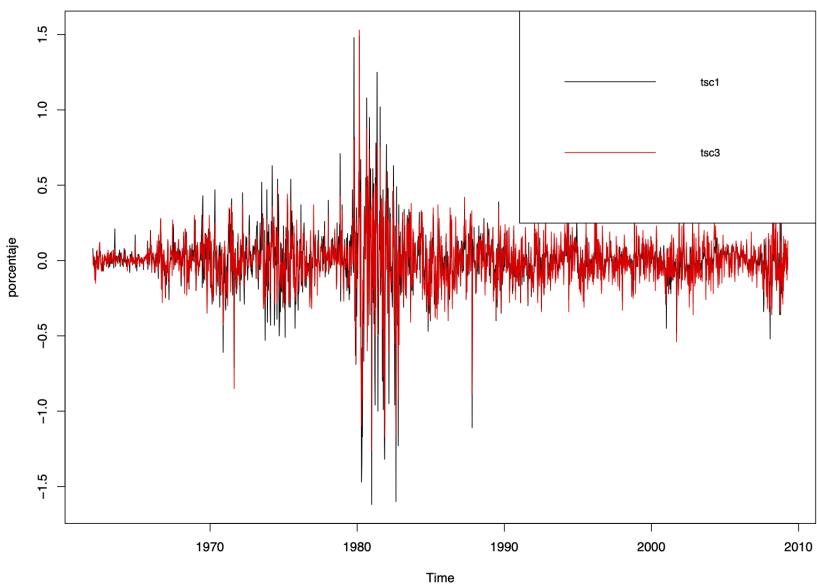


Figura 1.7: Gráficas de las variables en cambios

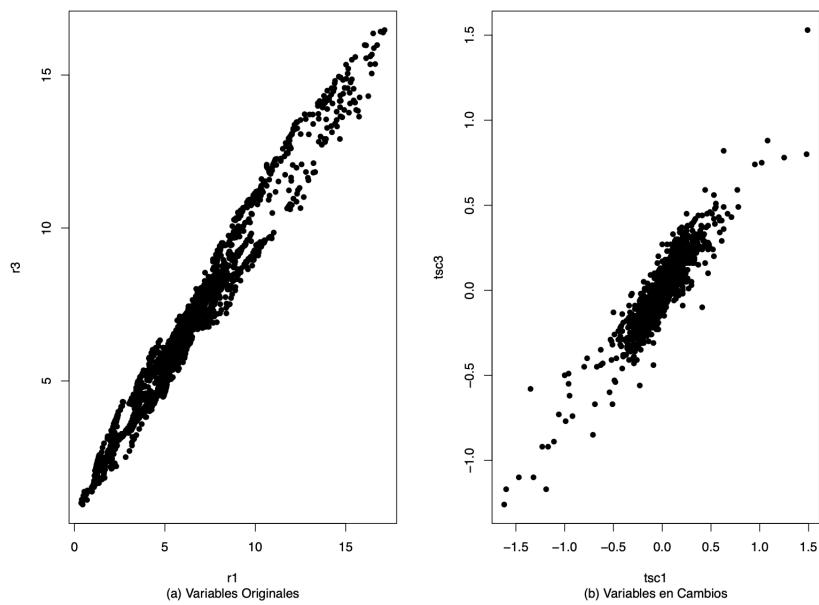


Figura 1.8: Gráficas de dispersión

2

Procesos Estocásticos y Procesos Lineales

En éste capítulo vamos a construir el marco teórico para el análisis de series de tiempo. Los modelos de series de tiempo están basados en procesos estocásticos, por lo tanto definiremos que es un proceso estocástico y daremos algunos resultados acerca de éstos. Después, introduciremos las ecuaciones en diferencias estocásticas como una forma de definir un modelo de series de tiempo. Más adelante veremos como los procesos estocásticos se relacionan con las ecuaciones en diferencias estocásticas o sistemas dinámicos discretos estocásticos.

Definición 2.1. *Una sucesión de variables aleatorias $\{X_t\}$ definidas sobre el mismo espacio de probabilidad (Ω, \mathcal{F}, P) es un proceso estocástico.*

Nota 2.2. *El índice t por lo general toma valores en un subconjunto \mathcal{T} de los números reales \mathbb{R} . Si \mathcal{T} es un conjunto discreto o enumerable, entonces diremos que el proceso estocástico es en tiempo discreto. Si \mathcal{T} es un conjunto no enumerable, entonces diremos que el proceso estocástico es en tiempo continuo.*

Ejemplos de conjuntos que pueden ser \mathcal{T} para el caso discreto son \mathbb{N} o \mathbb{Z} , y a $\{X_t\}$ se le conoce como un proceso estocástico en tiempo discreto. Si $\mathcal{T} = \mathbb{R}$ o $\mathcal{T} = [0, 1]$, entonces se le dice proceso estocástico en tiempo continuo. Para éste escrito, asumiremos que $\mathcal{T} = \mathbb{Z}$.

Nota 2.3. *Es frecuente que al proceso estocástico $\{X_t\}$ también se le llame serie de tiempo al igual que al conjunto de observaciones $\{x_t\}$. Por lo tanto, hay que tener claro en cada caso de que objeto se está hablando.*

Nota 2.4. Vale la pena decir que en este curso se trabajarán procesos estocásticos de valor real, es decir el espacio del estado es \mathbb{R} . Pero es importante precisar que los procesos estocásticos pueden tener como espacio del estado a \mathbb{R}^d , entonces hablaremos de series de tiempo multivariadas, o un espacio funcional como L^2 (espacio de funciones cuadrado integrables) y en este caso hablaremos de series de tiempo funcionales, o el espacio de los arreglos con entradas en los números reales, llamados tensores.

2.1 Procesos estocásticos Estacionarios

Desde el enfoque Estadístico del análisis de series de tiempo, son de vital importancia una familia de procesos estocásticos que son llamados estacionarios. Veremos de qué se trata. Hay dos definiciones de estacionariedad, y empezaremos hablando de estacionariedad en el sentido estricto.

Definición 2.5. Sea $\{X_t\}$ un proceso estocástico. Se dice que $\{X_t\}$ es estrictamente estacionario o estacionario en el sentido estricto si

$$(X_{t_1}, X_{t_2}, \dots, X_{t_n}) \stackrel{d}{=} (X_{t_1+h}, X_{t_2+h}, \dots, X_{t_n+h})$$

para todo $n \in \mathbb{N}^*$, todos los puntos $(t_1, t_2, \dots, t_n) \in \mathcal{T}^n$ y todos los saltos $h \in \mathbb{Z}$.

Lo que se entiende de la anterior definición es que la distribución de cualquier vector aleatorio finito dimensional que se puede formar con las variables del proceso $\{X_t\}$, tiene distribución invariante ante traslaciones en el tiempo.

Como ejemplo considere el proceso estocástico $\{X_t\}$ tal que las variables aleatorias son i.i.d., lo cual escribiremos en corto $\{X_t\} \sim i.i.d.$ Este proceso es trivialmente estacionario en el sentido estricto. En efecto, al ser i.i.d cada variable aleatoria unidimensional X_t tiene la misma distribución para cada t . De la misma manera, los vectores (X_{t_1}, X_{t_2}) y (X_{t_1+h}, X_{t_2+h}) para cada t_1, t_2 y h tienen la misma distribución dado que $\{X_t\}$ es i.i.d. El mismo razonamiento es válido para vectores de mayores dimensiones, por lo tanto $\{X_t\}$ es estrictamente estacionario.

Veamos un ejemplo no trivial:

Ejemplo 2.6. Sea $\{X_t\} \sim i.i.d.$, luego definimos el filtro lineal invariante finito de X_t como $Y_t = \sum_{j=-k}^k a_j X_{t-j}$ donde $(a_0, a_{\pm 1}, \dots, a_{\pm k}) \in \mathbb{R}^{2k+1}$, entonces el proceso $\{Y_t\}$ es estacionario en forma estricta.

Para chequear que $\{Y_t\}$ es estacionario en forma estricta, lo haremos basándonos en un caso particular para un vector aleatorio bidimensional, digamos (Y_{t_1}, Y_{t_2}) , para el cual encontraremos la función característica, y veremos que es igual que la función característica del vector aleatorio (Y_{t_1+h}, Y_{t_2+h}) .

Sea $Y_t = \sum_{j=-k}^k a_j X_{t-j}$ con $\{X_t\} \sim i.i.d.$ donde $(a_0, a_{\pm 1}, \dots, a_{\pm k}) \in \mathbb{R}^{2k+1}$. Recordemos como es la definición de la función característica y veamos como es la función característica para Y_t :

$$\begin{aligned}\varphi_{Y_t}(t) &= E[e^{itY_t}] \\ &= E\left[e^{it(\sum_{j=-k}^k a_j X_{t-j})}\right] \\ &= E\left[e^{\sum_{j=-k}^k ita_j X_{t-j}}\right] \\ &= E\left[\prod_{j=-k}^k e^{i(t a_j) X_{t-j}}\right] \\ &\stackrel{iid}{=} \prod_{j=-k}^k E\left[e^{i(t a_j) X_{t-j}}\right] \\ &= \prod_{j=-k}^k \varphi_{X_{t-j}}(t a_j)\end{aligned}$$

Para puntos en el tiempo t_1 y t_2 ,

$$\varphi_{Y_{t_1}}(v_1) = \prod_{j=-k}^k \varphi_{X_{t_1-j}}(v_1 a_j) \quad \text{y} \quad \varphi_{Y_{t_2}}(v_2) = \prod_{j=-k}^k \varphi_{X_{t_2-j}}(v_2 a_j)$$

Veamos ahora como es la función característica para el vector aleatorio

bidimensional $(Y_{t_1}, Y_{t_2})'$:

$$\begin{aligned}
 \varphi_{(Y_{t_1}, Y_{t_2})}(v_1, v_2) &= E \left[e^{i(v_1 Y_{t_1} + v_2 Y_{t_2})} \begin{pmatrix} Y_{t_1} \\ Y_{t_2} \end{pmatrix} \right] \\
 &= E \left[e^{i(v_1 Y_{t_1} + v_2 Y_{t_2})} \right] \\
 &= E \left[e^{i v_1 Y_{t_1} + i v_2 Y_{t_2}} \right] \\
 &= E \left[e^{i v_1 Y_{t_1}} \cdot e^{i v_2 Y_{t_2}} \right] \\
 &= E \left[e^{i v_1 (\sum_{j=-k}^k a_j X_{t_1-j})} e^{i v_2 (\sum_{j=-k}^k a_j X_{t_2-j})} \right] \\
 &= E \left[e^{\sum_{j=-k}^k i v_1 a_j X_{t_1-j}} e^{\sum_{j=-k}^k i v_2 a_j X_{t_2-j}} \right] \\
 &= E \left[\exp \left\{ i (v_1 a_k \quad \dots \quad v_1 a_0 \quad \dots \quad v_1 a_{-k}) \begin{pmatrix} X_{t_1-k} \\ \vdots \\ X_{t_1} \\ \vdots \\ X_{t_1+k} \end{pmatrix} \right\} \right. \\
 &\quad \left. \exp \left\{ i (v_2 a_k \quad \dots \quad v_2 a_0 \quad \dots \quad v_2 a_{-k}) \begin{pmatrix} X_{t_2-k} \\ \vdots \\ X_{t_2} \\ \vdots \\ X_{t_2+k} \end{pmatrix} \right\} \right] \\
 &= E \left[\exp \left\{ i \begin{pmatrix} v_1 a_k = \varepsilon_1 \\ \vdots \\ v_1 a_0 = \varepsilon_{k+1} \\ \vdots \\ v_1 a_{-k} = \varepsilon_{2k+1} \\ v_2 a_k = \varepsilon_{2k+2} \\ \vdots \\ v_2 a_0 = \varepsilon_{3k+2} \\ \vdots \\ v_2 a_{-k} = \varepsilon_{4k+2} \end{pmatrix}^T \begin{pmatrix} X_{t_1-k} \\ \vdots \\ X_{t_1} \\ \vdots \\ X_{t_1+k} \\ X_{t_2-k} \\ \vdots \\ X_{t_2} \\ \vdots \\ X_{t_2+k} \end{pmatrix} \right\} \right] \\
 &= \varphi_{(X_{t_1-k}, \dots, X_{t_1}, \dots, X_{t_1+k}, X_{t_2-k}, \dots, X_{t_2}, \dots, X_{t_2+k})}(\varepsilon_1, \dots, \varepsilon_{4k+2})
 \end{aligned}$$

Sin pérdida de generalidad asumiremos que no hay variables X'_t s repetidas en ambas sumas.

Se ha probado entonces que

$$\varphi_{Y_{t_1}, Y_{t_2}}(v_1, v_2) =$$

$$\varphi_{(X_{t_1-k}, \dots, X_{t_1}, \dots, X_{t_1+k}, X_{t_2-k}, \dots, X_{t_2}, \dots, X_{t_2+k})}(\varepsilon_1, \dots, \varepsilon_{4k+2})$$

$$\begin{aligned} & \varphi_{Y_{t_1+h}, Y_{t_2+h}}(v_1, v_2) = \\ & \varphi_{(X_{t_1+h-k}, \dots, X_{t_1+h-k}, \dots, X_{t_1+h-k}, X_{t_2+h-k}, \dots, X_{t_2+h}, \dots, X_{t_2+h+k})}(\varepsilon_1, \dots, \varepsilon_{4k+2}) \end{aligned}$$

Comparando $\varphi_{Y_{t_1}, Y_{t_2}}(v_1, v_2)$ con $\varphi_{Y_{t_1+h}, Y_{t_2+h}}(v_1, v_2)$, la *f.g.m.* $\varphi_{Y_{t_1}, Y_{t_2}}(\cdot, \cdot)$ depende de

$$(X_{t_1-k}, \dots, X_{t_1}, \dots, X_{t_1+k}, X_{t_2-k}, \dots, X_{t_2}, \dots, X_{t_2+k})$$

mientras que $\varphi_{Y_{t_1+h}, Y_{t_2+h}}(\cdot, \cdot)$ depende de

$$(X_{t_1-k+h}, \dots, X_{t_1-h}, \dots, X_{t_1+k+h}, X_{t_2-k+h}, \dots, X_{t_2+h}, \dots, X_{t_2+k+h})$$

Pero estos dos vectores tienen la misma distribución dado que $\{X_t\} \sim i.i.d.$
Esto termina la prueba.

Otra clase de estacionariedad que es menos restrictiva, es la estacionariedad débil o de segundo orden. La estacionariedad débil impone condiciones únicamente sobre los dos primeros momentos del proceso estocástico, es decir, sobre los valores esperados y sobre las covarianzas de proceso. Veamos ahora dos definiciones previas antes de definir la estacionariedad de segundo orden. A menos que se diga otra cosa, el espacio del estado de las variables aleatorias del proceso se asumirá que es \mathbb{R} .

Definición 2.7. *La función de medias del proceso estocástico $\{X_t\}$ se define como*

$$\mu_t = E[X_t]$$

asumiendo que el $E[X_t] < \infty$.

Definición 2.8. *Si $\{X_t\}$ es un proceso estocástico tal que $E|X_t|^2 < \infty$ para cada $t \in \mathbb{Z}$, entonces la función de autocovarianza(**facv**) $\gamma_X(\cdot, \cdot)$ de $\{X_t\}$ es definido como el segundo momento*

$$\gamma_X(r, s) = Cov(X_r, X_s) = E[(X_r - E[X_r])(X_s - E[X_s])]$$

para cada $r, s \in \mathbb{Z}$.

Note que la anterior definición nos muestra una forma de medir la asociación estadística lineal entre las variables aleatorias del proceso en los tiempos r y s . Por qué el nombre de Autocovarianza?

Definición 2.9. *El proceso estocástico $\{X_t\}$ es llamado estacionario(en el sentido débil o de segundo orden) si*

- (i) $E|X_t|^2 < \infty$ para todo t .

- (ii) $\mu_t = E[X_t] = m$ para todo t .
- (iii) $\gamma_X(r, s) = \gamma_X(r + t, s + t)$ para todo r, s, t .

Así para $t = -s$ en el numeral (iii), tenemos que $\gamma_X(r, s) = \gamma_X(r - s, s - s) = \gamma_X(r - s, 0) := \gamma(r - s)$, es decir, función $\gamma_X(,)$ depende sólo r, s a través de su diferencia $r - s = h$.

Nota 2.10. La condición (i) de la definición anterior se entiende que el segundo momento debe ser finito para que posteriormente podamos definir adecuadamente la correlación.

La condición (ii) se entiende que la función de medias es constante. Mientras que la condición (iii) establece que para que un proceso estocástico sea estacionario la asociación estadística debe depender únicamente de la diferencia entre los tiempos $h = r - s$, es decir no depende de los tiempos r, s . A la diferencia h se le conoce como rezago, y entonces diremos que $\gamma_X(h) = Cov(X_{t+h}, X_t)$ es la covarianza de las variables aleatorias en el rezago h , es decir, no depende del tiempo t en el escenario de estacionariedad de segundo orden.

Ejemplo 2.11. Considere un proceso estocástico $\{Z_t\}$. Diremos que el proceso es ruido blanco si

1. $E[Z_t] = 0$ para todo t .
2. La función de autocovarianza

$$\gamma(s, t) = \begin{cases} 0 & \text{si } s \neq t, \\ \sigma_Z^2 & \text{si } s = t. \end{cases}$$

Entonces trivialmente $\{Z_t\}$ es estacionario en el sentido débil, dado que por una parte sus valores esperados son constantes, y por otra parte sus covarianzas $Cov(X_{t+h}, X_t)$ o son cero o son σ_Z^2 , es decir no depende del tiempo t . Denotaremos al proceso como $\{Z_t\} \sim RB(0, \sigma_Z^2)$.

Ejemplo 2.12. Si $\{X_t\}$ es tal que las variables aleatorias son independientes y

$$X_t \sim \begin{cases} \exp(\lambda = 1), & t \text{ impar}, \\ N(\mu = 1, \sigma = 1), & t \text{ par}. \end{cases}$$

Así, $E[X_t] = 1$ y $Var[X_t] = 1$ para todo t .

$Cov(X_t, X_{t+h}) = 0$ para todo $h \geq 1$. Es decir, el proceso $\{X_t\}$ es estacionario

en forma débil o de segundo orden. Aunque $\{X_t\}$ no es estacionario en forma estricta porque por ejemplo, las distribuciones unidimensionales no son iguales para todo t .

Ejemplo 2.13. Sea $X_t = A \cos(\theta t) + B \sin(\theta t)$, $\theta \in (-\pi, \pi)$ y A, B variables aleatorias no correlacionadas con medias cero y varianzas 1. Computemos la función de autocovarianza.

$$\begin{aligned} Cov(X_{t+h}, X_t) &= Cov(A \cos(\theta(t+h)) + B \sin(\theta(t+h)), A \cos(\theta t) + B \sin(\theta t)) \\ &= Cov(A \cos(\theta(t+h)), A \cos(\theta t)) + Cov(A \cos(\theta(t+h)), B \sin(\theta t)) \\ &\quad + Cov(B \sin(\theta(t+h)), A \cos(\theta t)) + Cov(B \sin(\theta(t+h)), B \sin(\theta t)) \\ &= \cos(\theta(t+h)) \cos(\theta t) Cov(A, A) + \cos(\theta(t+h)) \sin(\theta t) Cov(A, B) \\ &\quad + \sin(\theta(t+h)) \cos(\theta t) Cov(B, A) + \sin(\theta(t+h)) \sin(\theta t) Cov(B, B) \\ &= \cos(\theta(t+h)) \cos(\theta t) \underbrace{Cov(A, A)}_0 + \cos(\theta(t+h)) \sin(\theta t) \underbrace{Cov(A, B)}_0 \\ &\quad + \sin(\theta(t+h)) \cos(\theta t) \underbrace{Cov(B, A)}_0 + \sin(\theta(t+h)) \sin(\theta t) \underbrace{Cov(B, B)}_1 \\ &= \cos(\theta(t+h)) \cos(\theta t) + \sin(\theta(t+h)) \sin(\theta t) \end{aligned}$$

Haciendo $\alpha = \theta(t+h)$ y $\beta = \theta t$, entonces $\alpha - \beta = \theta h$ y aplicando la identidad $\cos(\alpha - \beta) = \cos(\alpha) \cos(\beta) + \sin(\alpha) \sin(\beta)$ tenemos que $\cos(\theta(t+h)) \cos(\theta t) + \sin(\theta(t+h)) \sin(\theta t) = \cos(\theta h)$.

Por lo tanto, $\gamma_X(h) = \cos(\theta h)$. Es decir la función de autocovarianza no depende de t , únicamente de h .

Por otro lado,

$$E[X_t] = E[A \cos(\theta t) + B \sin(\theta t)] = \cos(\theta t) E[A] + \sin(\theta t) E[B] = 0.$$

Finalmente, $Var(X_t) = \gamma_X(0) = 1 < \infty$. Por lo tanto, el proceso $\{X_t\}$ es estacionario de segundo orden.

Ejemplo 2.14. Sea $\{S_t\}$ la caminata aleatoria definida como $S_t = X_1 + X_2 + \dots + X_t$ donde $\{X_t\} \sim \text{i.i.d}$ con media cero y varianza σ^2 , y definiendo $S_0 = 0$.

Note que $E[S_t] = E[X_1 + X_2 + \dots + X_t] = 0$.

Por otro lado, computemos $\gamma_S(t+h, t)$.

$$\begin{aligned} \gamma_S(t+h, t) &= Cov(S_{t+h}, S_t) = Cov(\underbrace{X_1 + X_2 + \dots + X_t}_{S_t} + X_{t+1} + \dots + X_{t+h}, S_t) \\ &= Cov(S_t + X_{t+1} + \dots + X_{t+h}, S_t) \\ &= Cov(S_t, S_t) + \underbrace{Cov(X_{t+1}, S_t)}_0 + \underbrace{Cov(X_{t+2}, S_t)}_0 + \dots + \underbrace{Cov(X_{t+h}, S_t)}_0 \end{aligned}$$

La razón por la cual algunas de las cantidades anteriores que son cero, es que por ejemplo X_{t+1} está en un tiempo adelante de S_t , lo cual las hace variables aleatorias independientes.

Así $\gamma_S(t+h, t) = Cov(S_t, S_t) = Var(S_t) = t\sigma^2$, debido a la independencia

de las variables aleatorias X_t . Lo anterior implica que la caminata aleatoria no es estacionaria en el sentido débil. Además, acá se puede ver que la varianza es creciente con el tiempo.

Nota 2.15. *Existe alguna relación entre los dos tipos de estacionariedad? Es decir alguna implica la otra? La respuesta es sí. Siempre la estacionariedad estricta implica la débil, asumiendo que tanto $E[X_t]$ como $E[X_t^2]$ sean finitas.*(Tarea)

Sin embargo, la otra implicación no es cierta, a menos que el proceso sea Gaussiano.

Vamos a establecer que cuando nos refiramos a un proceso estacionario, será en el sentido débil a menos que se diga otra cosa. Vale la pena decir, que cuando se diga que una serie es estacionaria, es por que hereda el nombre del proceso estocástico que lo generó, es decir, la serie de tiempo parece heredar las propiedades empíricas del proceso estocástico que lo generó.

Ejercicio 2.16. Realice los siguientes ejercicios:

- ✓ Realice el punto 1.1, 1.2(a. y b.)1.4, 1.5 ,1.8, 1.10 y 1.15 del libro [3].
- ✓ Realice el punto 1.11, 1.13 del libro [2].
- ✓ Considere el procesos estocástico

$$Z_t = \mu + R \sin(\omega t + \theta) + a_t,$$

donde ω es la frecuencia angular dada por $\omega = \frac{2\pi}{s}$ donde s es el periodo estacional, R es la amplitud y $\{a_t\} \sim RB(0, \sigma^2)$. Considere que $s = 4$, encuentre la función de medias y la función de autocovarianza del procesos $\{Z_t\}$. Qué puede usted observar?

- ✓ Sea X una variable aleatoria con distribución uniforme sobre $(0, \pi)$. Considere la sucesión de variables aleatorias $\{Y_t, t \in \mathbb{N}\}$ donde $Y_t = \cos(tX)$. Discuta las propiedades de la sucesión aleatoria.

Definición 2.17. Un proceso $\{X_t\}$ es llamado Gaussiano si los vectores n -dimensionales $\mathbf{X} = (X_{t_1}, X_{t_2}, \dots, X_{t_n})'$, para cada colección de puntos distintos t_1, t_2, \dots, t_n , y cada entero positivo n , tiene una distribución normal multivariada no singular.

Si un proceso $\{X_t\}$ es estacionario en el sentido débil y también es Gaussiano, entonces $\{X_t\}$ es estacionario en el sentido estricto.(Tarea.) Adicionalmente, hacer los ejercicios del libro de Brockwell and Davis Página 40

Ejercicios 1.7. Finalmente, encuentre la media y la facv del proceso definido en el ejemplo 2.6.

En seguida mostraremos algunos ejemplos simulados de series de tiempo que provienen de procesos estocásticos estacionarios. Por ejemplo, la figura 2.1 corresponde a la simulación de un proceso ARMA. Note que su comportamiento es estable, es decir su función de medias parece ser constante al igual que su varianza, lo cual nos hace pensar que el proceso es estacionario. En la figura 2.2, podemos observar la simulación de un proceso ARCH. Note que la serie de tiempo parece ser estacionaria. Note que en la figura 2.3 podemos observar un proceso estable en media, sin embargo en ocasiones aparecen valores bastante extremos, lo cual es natural dado que el proceso generador tiene distribución de Cuachy. Por lo tanto, la serie no sería estacionaria.

Ejercicio 2.18. *De las series de tiempo reales mencionadas anteriormente, cuales parecen provenir de procesos estocásticos estacionarios? Las series de empleo, desempleo, PIB son estacionarias?*

Nota 2.19. *En la práctica muchas series son estacionarias, sin embargo, es más común encontrarnos con series que son no estacionarias, y entonces por qué estudiar las series de tiempo estacionarias? Porque las metodologías para analizar series de tiempo estacionarias sugieren que primero la serie sea transformada a estacionaria, luego le sea ajustado un modelo a la serie estacionaria, y posteriormente se vayan incorporando las características de no estacionariedad al modelo.*

Vamos a estudiar algunas propiedades importantes que tiene la función de autocovarianza de un proceso estocástico estacionario. La facv de un proceso estocástico prácticamente lo caracteriza, por éste motivo se debe explorar un poco sus propiedades. Adicionalmente, la facv va a ser útil en la obtención del pronóstico.

Proposición 2.20. *Considere que tenemos un proceso estocástico estacionario $\{X_t\}$, y que $\gamma(h) = \text{Cov}(X_{t+h}, X_t) = E[(X_{t+h} - \mu)(X_t - \mu)]$, entonces $\gamma(\cdot)$ satisface:*

- i) $\gamma(0) \geq 0$
- ii) $|\gamma(h)| \leq \gamma(0)$ para todo $h \in \mathbb{Z}$

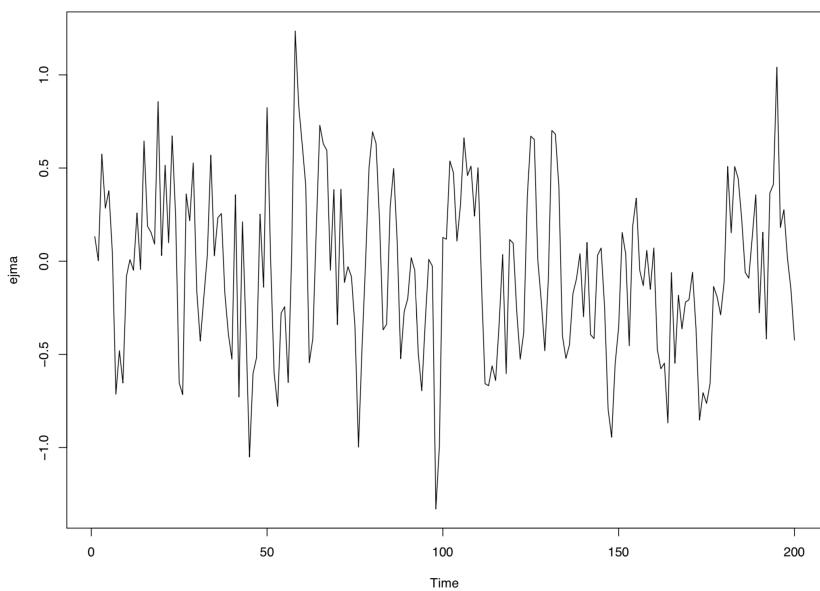


Figura 2.1: Serie de Tiempo Simulada de un proceso ARMA con errores Normales

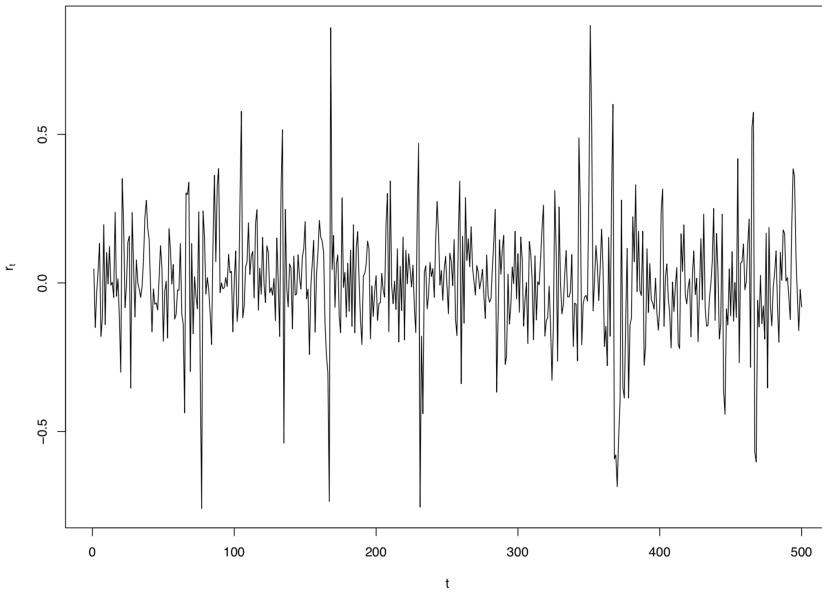


Figura 2.2: Serie de Tiempo Simulada de un proceso ARCH

iii) Es una función par, es decir,

$$\gamma(h) = \gamma(-h), \text{ para todo } h \in \mathbb{Z}$$

iv. $\gamma(h)$ es un función definida no negativa; es decir, para cualquier conjunto de constantes $(a_1, \dots, a_n) \in \mathbb{R}^n$, puntos del tiempo $(t_1, \dots, t_n) \in \mathbb{Z}^n$, y cualquier \mathbb{N}^* ,

$$\sum_{i=1}^n \sum_{j=1}^n a_i \gamma(t_i - t_j) a_j \geq 0.$$

Demostración. i) Como $\gamma(0) = \text{Cov}(X_t, X_t) = \text{Var}(X_t) \geq 0$.

ii) Por definición de la función de covarianza tenemos que $\gamma(h) = \text{Cov}(X_{t+h}, X_t) = E[(X_{t+h} - \mu)(X_t - \mu)]$.

Ahora, usando la desigualdad de Cauchy-Schwarz ($|E[XY]| \leq \sqrt{E[X^2]} \sqrt{E[Y^2]}$) tenemos que

$$\begin{aligned} |\gamma(h)| &= |E[(X_{t+h} - \mu)(X_t - \mu)]| \leq \sqrt{E[(X_{t+h} - \mu)^2]} E[(X_t - \mu)^2] \\ &= \sqrt{\text{Var}(X_t)} \sqrt{\text{Var}(X_t)} = \text{Var}(X_t) = \gamma(0). \end{aligned}$$

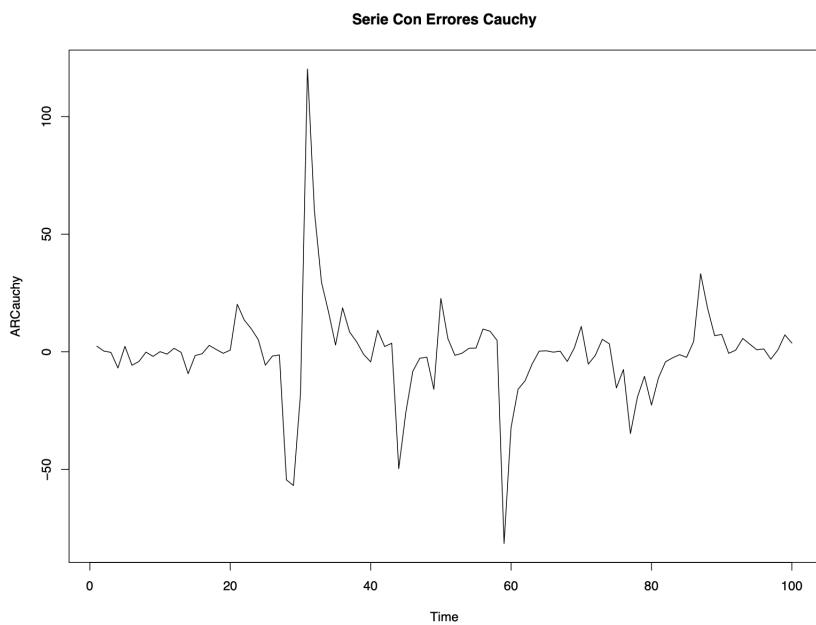


Figura 2.3: Serie de Tiempo Simulada de un proceso AR(1) con errores Cauchy

Con lo cual $|\gamma(h)| \leq \gamma(0)$.

iii) Ahora, $\gamma(-h) = Cov(X_{t-h}, X_t) = Cov(X_t, X_{t+h}) = Cov(X_{t+h}, X_t) = \gamma(h)$.

Lo anterior hace que usemos frecuentemente sólo rezagos positivos.

(iv) Note que para cualquier muestra $\{X_{t_1}, \dots, X_{t_n}\}$ de un proceso estocástico estacionario $\{X_t\}$, se tiene que $E|\sum_{i=1}^n a_i(X_{t_i} - \mu)|^2 \geq 0$, por lo tanto tenemos la desigualdad con apenas unos cálculos algebraicos. \square

Sigue de la anterior proposición que la matriz de covarianza, Γ_n , de una muestra de tamaño n , $\{X_{t_1}, \dots, X_{t_n}\}$, de un proceso estocástico estacionario es simétrica y definida no negativa, donde

$$\Gamma_n = Cov(X_{t_1}, \dots, X_{t_n}) = \begin{bmatrix} \gamma(0) & \gamma(t_1 - t_2) & \cdots & \gamma(t_1 - t_n) \\ \gamma(t_2 - t_1) & \gamma(0) & \cdots & \gamma(t_2 - t_n) \\ \vdots & \vdots & \ddots & \vdots \\ \gamma(t_n - t_1) & \gamma(t_n - t_2) & \cdots & \gamma(0) \end{bmatrix}.$$

Note que desde el punto de vista estadístico y práctico la correlación nos ofrece una manera mas conveniente de interpretar la asociación estadística lineal que existe entre dos variables, es por esto que vamos a entrar a definir la función de autocorrelación(facv) a través de la facv.

Definición 2.21. *La función de autocorrelación de un proceso estocástico se define como*

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

Nota 2.22. *Si $\{X_t\}$ es estacionario, entonces la anterior definición se reduce a*

$$\rho(h) = Cor(X_{t+h}, X_t) = \frac{\gamma(h)}{\gamma(0)}.$$

Si $\{X_t\}$ es un proceso estocástico estacionario, entonces la función de medias es $\mu_t = \mu$, así que con una sola muestra X_1, \dots, X_T es posible estimar a μ usando el estimador usual $\bar{X} = \frac{1}{T} \sum_{t=1}^T X_t$. El estimador de la facv se define como $\hat{\gamma}(h) = \frac{1}{T} \sum_{t=1}^{T-h} (X_{t+h} - \bar{X})(X_t - \bar{X})$. para $h = 0, 1, \dots, T - 1$. De manera análoga se puede definir la autocorrelación muestral como sigue:

$$\hat{\rho}(h) = \frac{\hat{\gamma}(h)}{\hat{\gamma}(0)}, h = 0, 1, \dots, T - 1$$

Ejemplos en R y Python para gráficas y el cálculo de facv y fac se pueden encontrar en los scripts *Importacion.Rmd* y *Importacion.ipynb*.

Tarea: Cual es función de autocorrelación de la caminata aleatoria? Compute $\rho(t, t - 1)$ y $\rho(t, t - 2)$, que puede usted decir?

Note que $\text{Corr}(S_t, S_{t-1}) = \frac{\text{Cov}(S_t, S_{t-1})}{\sqrt{S_t} \sqrt{S_{t-1}}}$, y del ejemplo 2.14, tenemos que $\text{Cov}(S_t, S_{t-1}) = \sigma^2(t - 1)$ y que $\text{Var}(S_t) = \sigma^2 t$, por lo tanto se puede verificar que

$$\text{Corr}(S_t, S_{t-1}) = \sqrt{\frac{t-1}{t}} \approx 1 - \frac{1}{2t}.$$

Ejemplo 2.23. Consideré el filtro lineal invariante finito $\{X_t\}$ basado en un proceso ruido blanco $\{Z_t\} \sim RB(0, \sigma_z^2)$, tal que $X_t = \sum_{j=-k}^k a_j Z_{t-j}$, donde $(a_0, a_{\pm 1}, \dots, a_{\pm k}) \in \mathbb{R}^{2k+1}$. Entonces esa bastante evidente verificar que $E[X_t] = 0$ y

$$\gamma(h) = \text{Cov}(X_{t+h}, X_t) = \sum_{i=-k}^k \sum_{j=-k}^k a_i a_j \text{Cov}(Z_{t+h-i}, Z_{t-j}) = \sigma_z^2 \sum_{|j+h| \leq k} a_j a_{j+h}.$$

En efecto, esto se obtiene igualando $t + h - i = t - j$, con lo cual $i = j + h$ y nos restringimos a una única suma. En efecto, como $i = j + h$, entonces $j + h$ toma valores entre $j + h = -k$ y $j + h = k$, puesto que i toma valores entre $-k$ y k , por lo tanto, $|j + h| \leq k$. Así, dejamos el subíndice j como está, mientras que le subíndice i lo reemplazamos por $j + h$. Note que j no puede tomar valores mas pequeños que $-k$, ni mas grandes que k . En general puede verificarse que si tomamos un proceso $\{X_t\}$ estacionario, y aplicamos un filtro lineal, obtenemos un proceso $\{Y_t\}$ estacionario tal que $Y_t = \sum_{j=-k}^k a_j X_{t-j}$.

Nota 2.24. Hay mas filtros que son importantes, por ejemplo el filtro diferencia $Y_t = X_t - X_{t-1}$. Otro ejemplo de filtro es el de ajuste estacional, es decir, $Y_t = \sum_{j=-6}^6 a_j X_{t-j}$ donde $a_j = 1/12$ para $j = 0, \pm 1, \dots, \pm 5$, y $a_{\pm 6} = 1/24$.

2.1.1 Procesos Lineales

La idea de filtro lineal puede expandirse a un filtro lineal infinito de un proceso estocástico estacionario, tal que se complete el espacio L^2 , es decir, el espacio de V.A.'s que tienen segundo momento finito.

Nota 2.25. (*El espacio $L^2(\Omega, \mathcal{F}, P)$*) Para mas detalles técnico ver [2] Capítulo 2.

Consideré un espacio de probabilidad (Ω, \mathcal{F}, P) y la colección C de todas las

variables aleatorias definidas sobre Ω que satisface

$$E[X^2] = \int_{\Omega} X^2(\omega)P(d\omega) < \infty.$$

Se puede verificar que C dotado de la multiplicación αX , con $\alpha \in \mathbb{R}$ y $X \in C$; y la suma $X + Y \in C$ es un espacio vectorial.

En efecto,

$$E[(\alpha X)^2] = \alpha^2 E[X^2] < \infty.$$

Ahora, $E[(X + Y)^2] \leq 2E[X^2] + 2E[Y^2]$ puesto que se puede verificar que $(X + Y)^2 \leq 2X^2 + 2Y^2$.

Con lo cual $E[(X + Y)^2] \leq 2E[X^2] + 2E[Y^2] < \infty$.

Lo anterior consistía en ver que la suma vectorial, y el producto de vector por escalar está bien definido. Ahora faltaría por verificar las siguientes propiedades:

Propiedad asociativa: $(X + Y) + Z = X + (Y + Z)$.

Propiedad Conmutativa: $X + Y = Y + X$.

Existencia del Elemento Neutro: La V.A. degenerada 0 tiene segundo momento finito, y satisface la condición $X + 0 = 0 + X = X$.

Existencia del elemento Inverso: Para cada $X \in C$, existe $-X \in C$ tal que $X - X = 0$.

Distributiva con respecto a la suma vectorial: Sea $\alpha \in \mathbb{R}$, y $X, Y \in C$, entonces $\alpha(X + Y) = \alpha X + \alpha Y$.

Distributiva con respecto a la suma escalar: Sean $\alpha, \beta \in \mathbb{R}$ y $x \in C$, entonces $(\alpha + \beta)x = \alpha x + \beta x$.

Ahora, se define le producto interno en C como:

$$\langle X, Y \rangle = E[XY].$$

Verificar las propiedades del producto interno. Tarea

Todas se cumplen menos la propiedad que $\langle X, X \rangle = 0$ si y sólo si $X = 0$, por qué?

Por lo tanto se hace necesario definir las clases de equivalencias en C , es decir, dos elementos $X, Y \in C$ son equivalentes, es decir $X \sim Y$ si y sólo si

$$P[X = Y] = 1.$$

Cómo se definen las clases de equivalencia?

$$\{Y \in C : X \sim Y\} = [X].$$

No es difícil verificar ahora que las clases de equivalencia si forma ahora un espacio vectorial con producto interno. A éste conjunto de clases residuales lo llamaremos L^2 .

Nota 2.26. Por abuso de notación, se seguirá usando X en vez de $[X]$. Al tener un producto interno, automáticamente tenemos una norma $\|\cdot\|$. El espacio L^2 , o mas bien el espacio de clases de equivalencia es completo, es decir, toda sucesión de Cauchy es convergente.

Una sucesión $\{X_n\}$ de elementos de C es de Cauchy si $\|X_n - X_m\| \rightarrow 0$ cuando $m, n \rightarrow \infty$.

Es decir, el espacio L^2 es completo, con lo cual lo llamaremos espacio de Hilbert. Al dotar al espacio de variables aleatorias con segundo momento finito C de una estructura de espacio de Hilbert, es posible usar todas sus propiedades para obtener algunos resultados. Por ejemplo, podemos usar el teorema de la proyección, obtener criterios para sucesiones de variables aleatorias, etc.

Proposición 2.27. Considere un filtro lineal invariante en el tiempo definido como la convolución de la forma

$$Y_t = \sum_{j=-\infty}^{\infty} a_j X_{t-j}, \quad \sum_{j=-\infty}^{\infty} |a_j| < \infty, \quad (2.1)$$

para cada $t \in \mathbb{Z}$. Si $\{X_t\}$ es una sucesión de variables aleatorias tal que $\sup_{t \in \mathbb{Z}} E[|X_t|] < \infty$, entonces la serie es absolutamente convergente P.c.s. Si, adicionalmente, $\sup_{t \in \mathbb{Z}} E[|X_t|^2] < \infty$, la serie converge en media cuadrática al mismo límite. En particular, si $\{X_t\}$ es estacionario, esas propiedades se cumplen. Sacado del libro [8].

Nota 2.28. A la condición $\sum_{j=-\infty}^{\infty} |a_j| < \infty$ se le conoce como que la serie de valores reales $\{a_j\}$ sea absolutamente convergente.

Vamos a utilizar el teorema de Riesz-Fisher para elementos en L^2 que dice:

Sea $\{U_n : n \in \mathbb{N}\}$ una sucesión de V.A.'s en L^2 . Entonces existe un elemento $U \in L^2$ tal que $U_n \xrightarrow{m.c.} U$ si y sólo si

$$\lim_{m \rightarrow \infty} \sup_{n \geq m} E|U_n - U_m|^2 = 0 \quad \rightarrow \quad \text{Probar que es de Cauchy}$$

Teorema de complez de L^2 : $\|U_n - U_m\|^2 = E|U_n - U_m|^2$.

Vamos a probar primero que la serie $\sum_{j=-\infty}^{\infty} a_j X_{t-j}$ es convergente en M.C.
Definamos el filtro lineal finito como

$$Y_t^n = \sum_{j=-n}^n a_j X_{t-j} \quad n \in \mathbb{N}$$

y entonces la convergencia en M.C. se hace sobre la sucesión $\{Y_t^n : n \in \mathbb{N}\}$ para cada $t \in \mathbb{Z}$. Entonces usaremos el teorema de Riesz-Fisher para la prueba. Por lo tanto sería suficiente probar que

$$\lim_{m \rightarrow \infty} \sup_{n \geq m} E|Y_t^n - Y_t^m|^2 = 0$$

para probar que $\{Y_t^n : n \in \mathbb{N}\}$ es convergente en M.C.

Para $n > m > 0$,

$$\begin{aligned} 0 &\leq E |Y_t^n - Y_t^m|^2 \\ &= E \left| \sum_{j=-n}^n a_j X_{t-j} - \sum_{j=-m}^m a_j X_{t-j} \right|^2 \quad \text{De la definición de } Y_t^n \\ &= E \left| \sum_{m < |j| \leq n} a_j X_{t-j} \right|^2 \quad (\star \text{ y ya que } n > m > 0) \\ &= E \left[\sum_{m < |j| \leq n} \sum_{m < |k| < n} a_j a_k X_{t-j} X_{t-k} \right] \quad (\sum_{i=1}^N a_i)^2 = \sum_{i=1}^N \sum_{j=1}^N a_i a_j \\ &= \left| \sum_{m < |j| \leq n} \sum_{m < |k| \leq n} a_j a_k E[X_{t-j} X_{t-k}] \right| \quad \text{Dado que toda la expresión es positiva} \\ &\leq \sum \sum |a_j| |a_k| |E[X_{t-j} X_{t-k}]| \quad \text{des. triangular} \\ &\leq \sum \sum |a_j| |a_k| (E [|X_{t-j}|^2])^{1/2} (E [|X_{t-k}|^2])^{1/2} \quad \text{Cauchy-Schwarz} \end{aligned}$$

Note que se puede ver un caso particular para chequear (\star) asumiendo $n = 3, m = 2$ dentro del valor esperado, tenemos que

$$\sum_{j=-3}^3 a_j X_{t-j} - \sum_{j=-2}^2 a_j X_{t-j} =$$

$$a_{-3} X_{t+3} + a_{-2} X_{t+2} + a_{-1} X_{t+1} + a_0 X_t + a_1 X_{t-1} + a_2 X_{t-2} + a_3 X_{t-3} -$$

$$a_{-2} X_{t+2} + a_{-1} X_{t+1} + a_0 X_t + a_1 X_{t-1} + a_2 X_{t-2}$$

=

$$a_{-3}X_{t+3} + a_3X_{t-3}.$$

Note que quedan las colas, es decir, los valores mas grandes que m y mas pequeños o iguales a n .

Retomando, dado que $\sup_{t \in \mathbb{Z}} E[|X_t|^2] < \infty$ y por la definición del sup, se tiene que para cada $t \in \mathbb{Z}$, $E[|X_t|^2] \leq \sup_{t \in \mathbb{Z}} E[|X_t|^2]$ entonces tenemos que

$$\begin{aligned} & \sum \sum |a_j||a_k| (E [|X_{t-j}|^2])^{1/2} (E [|X_{t-k}|^2])^{1/2} \\ & \leq \sum \sum |a_j||a_k| \left(\sup_{t \in \mathbb{Z}} E[|X_t|^2] \right)^{1/2} \left(\sup_{t \in \mathbb{Z}} E[|X_t|^2] \right)^{1/2} \\ & = \left(\sup_{t \in \mathbb{Z}} E[|X_t|^2] \right) \sum \sum |a_j||a_k| \\ & = \underbrace{\sup_{t \in \mathbb{Z}} E[|X_t|^2]}_{<\infty} \left(\sum_{m < |j| \leq n} |a_j| \right)^2 \end{aligned}$$

Cuando $m, n \rightarrow \infty$ entonces $\sum_{m < |j| \leq n} |a_j| \rightarrow 0$ ya que $\{a_j\}$ es absolutamente convergente. Entonces

$$E[|Y_t^n - Y_t^m|^2] = ||Y_t^n - Y_t^m||^2 \xrightarrow{n,m \rightarrow \infty} 0$$

Es decir, la serie es de Cauchy, por lo tanto es convergente.

Note que hasta ahora se ha comprobado que la serie es convergente, pero no se ha indicado cuál es dicho límite. Veamos como se comprueba.

Sea \tilde{Y}_t el límite en M.C. de Y_t^n . Ahora usaremos el lema de Fatou para

establecer que \tilde{Y}_t y Y_t son iguales.

$$\begin{aligned}
 0 &\leq E \left[\left| \tilde{Y}_t - Y_t \right|^2 \right] \\
 &= E \left[\left| \tilde{Y}_t - \sum_{j=-\infty}^{\infty} a_j X_{t-j} \right|^2 \right] \\
 &= E \left[\left| \tilde{Y}_t - \lim_{n \rightarrow \infty} \sum_{j=-n}^n a_j X_{t-j} \right|^2 \right] \\
 &= E \left[\lim_{n \rightarrow \infty} \left| \tilde{Y}_t - \sum_{j=-n}^n a_j X_{t-j} \right|^2 \right] && \text{continuidad límite} \\
 &= E \left[\liminf_{n \rightarrow \infty} \underbrace{\left| \tilde{Y}_t - \sum_{j=-n}^n a_j X_{t-j} \right|^2}_{\text{V.A.'s } \geq 0} \right] \\
 &\leq \liminf_{n \rightarrow \infty} E \left| \tilde{Y}_t - \sum_{j=-n}^n a_j X_{t-j} \right|^2 && \text{Lema de Fatou} \\
 &= \liminf_{n \rightarrow \infty} E \left| \tilde{Y}_t - Y_t^n \right|^2 = 0
 \end{aligned}$$

Ya que \tilde{Y}_t es el límite en M.C. Entonces se ha probado que $E \left[\left| \tilde{Y}_t - Y_t \right|^2 \right] = 0$, así $\tilde{Y}_t = Y_t$ en M.C., es decir, Y_t es el límite en M.C. de Y_t^n . Ahora verifiquemos la convergencia casi segura.

Convergencia C.S.

Sea $\{X_n\}$ una sucesión de V.A.'s, diremos que $X_n \xrightarrow{c.s.} X$ si existe un evento $N \in \mathcal{B}$, tal que $P(N) = 0$ y si $w \in N^c$ entonces

$$\lim_{n \rightarrow \infty} X_n(w) \text{ existe.}$$

Para probar la convergencia *c.s.* usaremos el siguiente lema (pg. 31 introduction to statistical time series - Fuller 1996)

Lema 2.29. *Sea $\{Z_j\}$ una sucesión de V.A.'s definidas sobre el mismo espacio de probabilidad (Ω, \mathcal{F}, P) . Asumamos que*

$$\sum_{j=-\infty}^{\infty} E[|Z_j|] < \infty$$

entonces $\sum_{j=-\infty}^{\infty} Z_j$ converge c.s. y así $E\left[\sum_{j=-\infty}^{\infty} Z_j\right] = \sum_{j=-\infty}^{\infty} E[Z_j]$

¡Note que puede ser más fácil chequear la convergencia de una serie de números reales!

Como se desea verificar que $\sum_{j=-\infty}^{\infty} a_j X_{t-j}$ es convergente c.s., entonces usando el lema anterior, debemos chequear que $\sum_{j=-\infty}^{\infty} E[|a_j X_{t-j}|]$ sea finito.

$$\begin{aligned} \sum_{j=-\infty}^{\infty} E[|a_j X_{t-j}|] &\leq \sum_{j=-\infty}^{\infty} |a_j| \sup_{t \in \mathbb{Z}} E[|X_t|] \\ &= \underbrace{\sup_{t \in \mathbb{Z}} E[|X_t|]}_{<\infty} \underbrace{\sum_{j=-\infty}^{\infty} |a_j|}_{<\infty} < \infty \end{aligned}$$

Es decir, es finito. Así, $\sum_{j=-\infty}^{\infty} a_j X_{t-j}$ converge c.s. a Y_t .

Lo anterior muestra que el proceso estocástico definido en 2.1, está bien definido, es decir, la series con convergente tanto en Media Cuadrática como Casi seguramente P , bajo las condiciones de la proposición 2.27. De manera directa, entonces también es procesos estocástico es convergente en probabilidad.

Si al proceso $\{X_t\}$ suponemos adicionalmente que es estacionario, entonces el proceso filtrado $\{Y_t\}$ también es estacionario. Para ver esto, establecemos la siguiente proposición.

Proposición 2.30. *Si $\{X_t\}$ es un proceso estocástico estacionario con función de autocovarianza $\gamma_X(\cdot)$ y si $\sum_{j=-\infty}^{\infty} |\psi_j| < \infty$, entonces para cada t , la serie*

$$Y_t = \sum_{j=-\infty}^{\infty} \psi_j X_{t-j} = \sum_{j=-\infty}^{\infty} \psi_j B^j X_t = \psi(B) X_t$$

converge casi seguramente P y en media cuadrática al mismo límite. Adicionalmente, el proceso $\{Y_t\}$ es estacionario y la función de autocovarianza está dada por

$$\gamma_Y(h) = \sum_{j,k=-\infty}^{\infty} \psi_j \psi_k \gamma_X(h - j + k).$$

Observación: El operador B^j es el operador de retardo, y se define como $B^0 X_t = X_t$, $B^1 X_t = X_{t-1}$, $B^j X_t = X_{t-j}$. Mientras que el operador de adelanto se define como $B^{-j} X_t = X_{t+j}$.

Demostración. Note que de la estacionariedad del proceso $\{X_t\}$ se tiene que $\{Y_t\}$ es estacionario debido a la proposición 2.27.

Para probar que $\{Y_t\}$ es convergente *c.s.* y en M.C. basta con verificar que $E[|X_t|] \leq c_1$ y $E[|X_t|^2] \leq c_2$. Basados en la proposición 2.27, para cada $t \in \mathbb{Z}$, en efecto, $\{E[|X_t|]\}^2 \leq E[|X_t|^2] = c_2$, ya que $\{X_t\}$ es estacionario. Así que para cada t ,

$$E[|X_t|] \leq \sqrt{c_2} = c_1 \quad \text{y} \quad E[|X_t|^2] = c_2$$

Así que las condiciones de la proposición 2.27 se satisfacen, y de esta forma, $\{Y_t\}$ es convergente *c.s.* y en M.C.

Ahora vamos que es estacionario. Para esto, computaremos $E[Y_t]$. Basados en la proposición 2.7.1 del libro [3] acerca de las propiedades de sucesiones convergentes en M.C. , es decir

$$\lim_{n \rightarrow \infty} E[X_n] = \lim_{n \rightarrow \infty} \langle X_n, 1 \rangle = \langle \lim_{n \rightarrow \infty} X_n, 1 \rangle = \langle X, 1 \rangle = E[X] = E[\lim_{n \rightarrow \infty} X_n]$$

tenemos que

$$E[Y_t] = E \left[\sum_{j=-\infty}^{\infty} \psi_j X_{t-j} \right] = E \left[\lim_{n \rightarrow \infty} \underbrace{\sum_{j=-n}^n \psi_j X_{t-j}}_{Y_t^n} \right]$$

como se mostró que $\{Y_t^n\}$ es convergente en M.C., tenemos que

$$\begin{aligned} E \left[\lim_{n \rightarrow \infty} \sum_{j=-n}^n \psi_j X_{t-j} \right] &= \lim_{n \rightarrow \infty} E \left[\sum_{j=-n}^n \psi_j X_{t-j} \right] && \text{prop. 2.7.1 (i) de [2]} \\ &= \lim_{n \rightarrow \infty} \sum_{j=-n}^n \psi_j \underbrace{E[X_{t-j}]}_{\text{fijo } \forall t} && \text{no depende de } t \\ &= \lim_{n \rightarrow \infty} \sum_{j=-n}^n \psi_j E[X_t] \\ &= E[X_t] \lim_{n \rightarrow \infty} \sum_{j=-n}^n \psi_j \\ &= E[X_t] \underbrace{\sum_{j=-\infty}^{\infty} \psi_j}_{<\infty} && \{\psi_j\} \text{ es abs. conv.} \end{aligned}$$

Ahora, computemos $E[Y_{t+h}Y_t]$ antes de hallar $\gamma_Y(h)$.

$$\begin{aligned}
 E[Y_{t+h}Y_t] &= E \left[\left(\sum_{j=-\infty}^{\infty} \psi_j X_{t+h-j} \right) \left(\sum_{k=-\infty}^{\infty} \psi_k X_{t-k} \right) \right] \\
 &= E \left[\lim_{n \rightarrow \infty} \left(\sum_{j=-n}^n \psi_j X_{t+h-j} \right) \left(\sum_{k=-n}^n \psi_k X_{t-k} \right) \right] \\
 &\stackrel{2.7.1(iii)}{=} \lim_{n \rightarrow \infty} E \left[\left(\underbrace{\sum_{j=-n}^n \psi_j X_{t+h-j}}_{Y_{t+h}^n} \right) \left(\underbrace{\sum_{k=-n}^n \psi_k X_{t-k}}_{Y_t^n} \right) \right] \\
 &= \lim_{n \rightarrow \infty} \left[\sum_{j=-n}^n \sum_{k=-n}^n \psi_j \psi_k X_{t+h-j} X_{t-k} \right] \\
 &= \lim_{n \rightarrow \infty} \underbrace{\sum_{j=-n}^n \sum_{k=-n}^n \psi_j \psi_k}_{\text{!Hay que verificar que es convergente!}} E[X_{t+h-j} X_{t-k}]
 \end{aligned}$$

Se puede verificar que es convergente si se verifica que $\sum_{j,k=-n}^n |\psi_j \psi_k E[X_{t+h-j} X_{t-k}]|$

es convergente. Veamos:

$$\begin{aligned}
& \lim_{n \rightarrow \infty} \sum_{j=-n}^n \sum_{k=-n}^n |\psi_j \psi_k E[X_{t+h-j} X_{t-k}]| \\
&= \lim_{n \rightarrow \infty} \sum_{j=-n}^n \sum_{k=-n}^n |\psi_j \psi_k| |E[X_{t+h-j} X_{t-k}]| \\
&= \lim_{n \rightarrow \infty} \sum_{j=-n}^n \sum_{k=-n}^n |\psi_j \psi_k| |E[X_{t+h-j} X_{t-k}]| \\
&\leq \lim_{n \rightarrow \infty} \sum_{j=-n}^n \sum_{k=-n}^n |\psi_j \psi_k| (E|X_{t+h-j}|^2)^{1/2} (E|X_{t-k}|^2)^{1/2} \quad \text{Cauchy-Schwarz} \\
&= \lim_{n \rightarrow \infty} \sum_{j=-n}^n \sum_{k=-n}^n |\psi_j \psi_k| (E|X_t|^2)^{1/2} (E|X_t|^2)^{1/2} \quad \text{por la estacionariedad } \{X_t\} \\
&= \lim_{n \rightarrow \infty} \sum_{j=-n}^n \sum_{k=-n}^n |\psi_j \psi_k| \underbrace{E|X_t|^2}_{<\infty} \\
&= \lim_{n \rightarrow \infty} E|X_t|^2 \left(\sum_{j=-n}^n |\psi_j| \right) \left(\sum_{k=-n}^n |\psi_k| \right) \\
&= E|X_t|^2 \lim_{n \rightarrow \infty} \left(\underbrace{\sum_{j=-n}^n |\psi_j|}_{<\infty} \right) \left(\underbrace{\sum_{k=-n}^n |\psi_k|}_{<\infty} \right) < \infty
\end{aligned}$$

Así

$$\begin{aligned}
E[Y_{t+h} Y_t] &= \lim_{n \rightarrow \infty} \sum_{j,k=-n}^n \psi_j \psi_k E[X_{t+h-j} X_{t-k}] < \infty \\
&= \sum_{j,k=-\infty}^{\infty} \psi_j \psi_k E[X_{t+h-j} X_{t-k}] \\
&= \sum_{j,k=-\infty}^{\infty} \psi_j \psi_k \left\{ \gamma_X(h-j+k) + \underbrace{(E[X_t])^2}_{\text{no depende de } t} \right\}
\end{aligned}$$

no depende de t . Entonces

$$\gamma_Y(h) = E[Y_{t+h} Y_t] - E[Y_{t+h}] E[Y_t]$$

Haciendo aritmética se puede ver que

$$\gamma_Y(h) = \sum_{j,k=-\infty}^{\infty} \psi_j \psi_k \gamma_X(h-j+k)$$

□

Un caso particular del filtro lineal infinito invariante es considerar que el proceso $\{X_t\}$ sea en forma particular un proceso ruido blanco $\{Z_t\}$, a éste proceso lo llamaremos *proceso lineal*.

Definición 2.31. *Un proceso estocástico $\{X_t\}$ es llamado **proceso lineal** si*

$$X_t = \sum_{j=-\infty}^{\infty} \psi_j Z_{t-j}, \quad (2.2)$$

donde $\{Z_t\} \sim RB(0, \sigma_z^2)$ y $\{\psi_j\}$ es una sucesión absolutamente sumable.

Nota 2.32. *El **proceso lineal** definido en 2.2 es estacionario debido a la proposición 2.30 ya que el proceso $\{Z_t\}$ es estacionario. Más específicamente la media del proceso es $E[X_t] = 0$, y la función de autocovarianza está dada por*

$$\gamma_X(h) = \sigma_z^2 \sum_{j=-\infty}^{\infty} \psi_j \psi_{j+h}. \quad (2.3)$$

En la ecuación 2.2 puede incluirse una constante tal que

$$X_t = \mu + \sum_{j=-\infty}^{\infty} \psi_j Z_{t-j},$$

así, el proceso $\{X_t\}$ tiene media μ , y su covarianza no se ve afectada.

Nota 2.33. *Un proceso lineal $\{X_t\}$ está completamente caracterizado por su media μ y su función de autocovarianza $\gamma(h)$ o autocorrelación $\rho(h)$ puesto que todas las relaciones entre las variables del proceso son lineales. Sin embargo, un proceso lineal visto como un modelo estadístico para los datos $\{x_1, \dots, x_T\}$ es inconveniente, por qué?*

El proceso lineal será el proceso base sobre el cual girará la familia de procesos **ARMA** de la siguiente sección, es decir, un procesos ARMA se puede escribir como un proceso lineal.

Por otra parte, podemos ver que si adicionalmente en la definición 2.2 asumimos que el proceso $\{Z_t\}$ es Gaussiano, entonces el proceso $\{X_t\}$ es Gaussiano también. Ver [21] páginas 5 y 6.

2.2 Análisis Descriptivo de una Series de Tiempo

Lo primero que vamos a llevar a cabo es la importación de datos, después la creación de objetos de series de tiempo y finalmente como hacer un análisis exploratorio acerca de una serie de tiempo en cuestión.

2.2.1 Importación de datos en R y Python

La importación de la base de datos y la posterior creación de un objetos de series de tiempo va a ser importante para su subsecuente análisis. En los archivos *Importacion.ipynb* y *Importacion.Rmd* se encuentran los códigos de *Python* y *R* para la importación de bases de datos y la creación de objetos de objetos de series de tiempo. No me detendré mucho tiempo en esta parte de importación, sin embargo una lectura mas profunda puede hacerse usando el libro [24] en los capítulos 3 y 4 para *R*, y para el caso de *Python* puede leerse los libros [28] y [26].

2.2.2 Gráfico de una Serie de Tiempo

Generalmente un gráfico de la serie de tiempo nos puede mostrar varias características de una serie de tiempo, de las cuales ya hablamos en la introducción(*Tendencia, estacionalidad, ciclos, heterocedasticidad, etc.*). En ocasiones es claro al solo inspeccionar la gráfica de la serie de tiempo que dicha serie proviene de un proceso estocástico que no es estacionario. Veamos por ejemplo la serie de desempleo de los Estados Unidos 2.4 y de Colombia 2.5. Estas dos series mensuales presentan tendencia, estacionalidad, posibles ciclos y quizá varianza no constante.

Note que en la sección anterior, definimos los estimadores de la media y de la función de autocorrelación(es la dependencia entre los valores de la serie), lo cual con datos de series de tiempo es importante medir; así que debemos al menos poder estimar las autocorrelaciones con precisión. Sería difícil medir esa dependencia si la estructura de la dependencia no es regular o cambia en cada momento. Por lo tanto, para lograr cualquier análisis estadístico significativo de los datos de series de tiempo, será crucial que, al menos, las funciones media y autocovarianza satisfagan las condiciones de estacionariedad establecidas en la Definición 2.9. A menudo, este no es el caso, y en la siguientes secciones mencionaremos algunos métodos para :(a)minimizar los efectos de la no estacionariedad para que se puedan estudiar las propiedades estacionarias de la serie, sacado de [34]; (b) detectar y extraer la tendencia, estacionalidad y posibles ciclos que puedan presentar las series. Usaremos los archivos *Descriptivo.Rmd* y *Análisis Descriptivo de una Serie de Tiempo en Python.ipynb* para desarrollar esta sección.

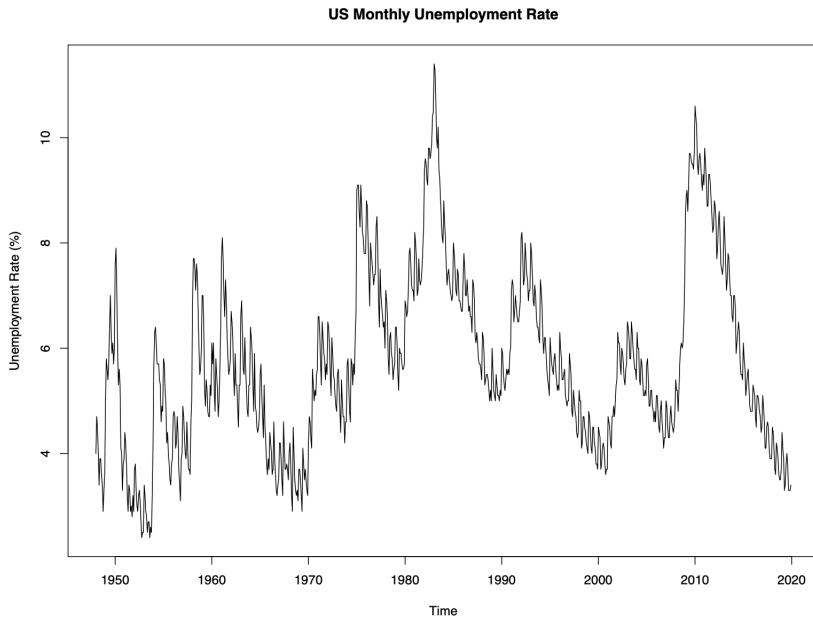


Figura 2.4: Tasa de Desempleo Estados Unidos

2.2.3 Análisis de Tendencias

Vamos a suponer inicialmente que nuestra serie observada $\{y_t\}$ presenta únicamente una tendencia determinista del tiempo, es decir

$$Y_t = \mu_t + a_t, \quad (2.4)$$

donde $\mu_t = f(t, \beta)$ es una función conocida y a_t es una componente aleatoria estacionaria(o no autocorrelacionado). Usualmente a la tendencia también se le conoce como nivel de la serie. Por lo general la tendencia se puede asumir polinómica, es decir

$$\mu_t = \beta_0 + \beta_1 t + \cdots + \beta_r t^r.$$

Sin embargo no hay restricción acerca de la forma de μ_t , sólo debería ser lineal en los parámetros para poder usar el método de mínimos cuadrados para la estimación, es decir, minimizar con respecto a los parámetros a:

$$\sum_{t=1}^T (Y_t - \mu_t)^2.$$

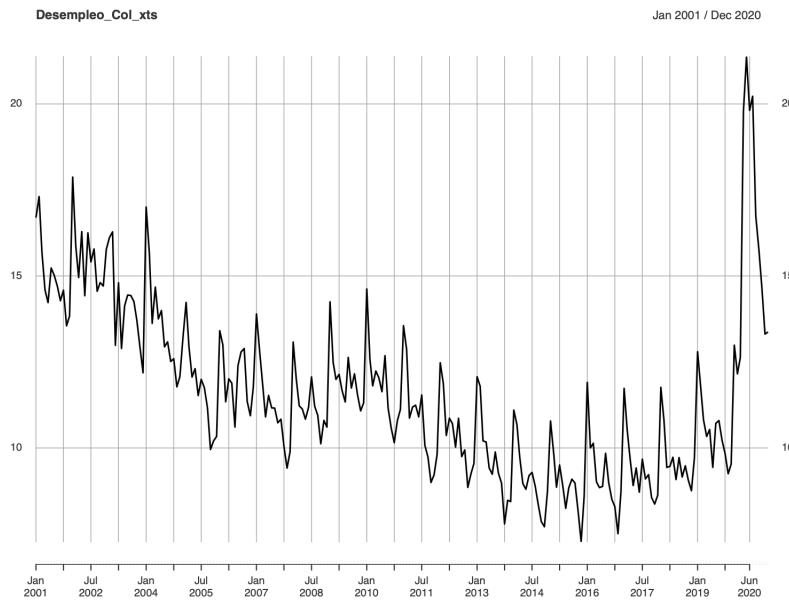


Figura 2.5: Tasa de Desempleo Colombia

Nota 2.34. ✓ También se puede considerar que la tendencia evolucione linealmente en el tiempo, es decir, nos permitiremos que la pendiente puede ser distinta en el tiempo como sigue:

$$\mu_t = \mu_{t-1} + \beta_{t-1}.$$

✓ Por otro lado, podemos ver que también que la tendencia no necesariamente puede ser determinista sino estocástica, es decir, una caminata aleatoria con drift(con constante)

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

puede ser un buen modelo para series que presentan tendencia, veamos una simulación de esta caminata aleatoria en la gráfica 6.6 donde $\{w_t\}$ es un proceso I.I.D.

Ejemplo 2.35. Consideraremos los datos del precio mensual(en centavos de dólar) del pollo por libra en los Estados Unidos que se encuentra en los datos chicken del paquete astsa. La gráfica de la serie de tiempo se muestra en la

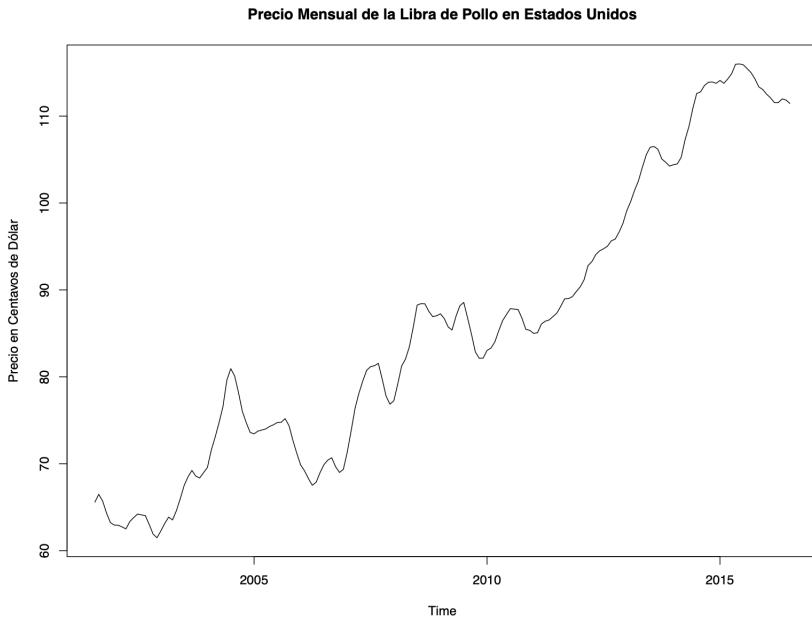


Figura 2.6: Precio del Pollo US Mensual

figura 2.6. La serie parece tener una tendencia lineal creciente, es decir, la podemos modelar como sigue:

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

Ahora, procedemos a estimar los parámetros de la tendencia determinista, con la cual obtenemos:

$\hat{\beta}_0 = -7131$ y $\hat{\beta}_1 = 3.592$, ambos altamente significativos. Ahora podemos ver la serie sin tendencia, ver Gráfica 2.7

Note que la serie sin tendencia no necesariamente es estacionaria.

Ejercicio 2.36. Repetir la experiencia con la serie de pasajeros encontrada en *data(AirPassengers)*.

Si se considera una caminata aleatoria como un modelo para la tendencia, uno puede ver que al diferenciar la serie $\{y_t\}$ de forma una ordinaria una sola vez($y_t - y_{t-1}$), se elimina dicha tendencia y podemos obtener una serie estacionaria.

Por supuesto al diferenciar una serie cuya tendencia es estocástica, podemos ver que obtenemos una serie estacionaria, mientras que si quitamos la

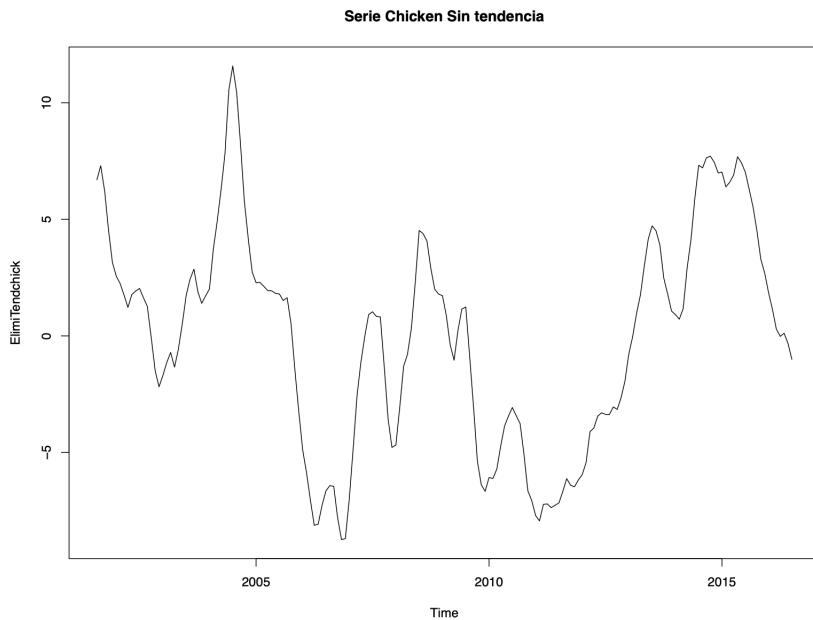


Figura 2.7: Serie Precio Pollo Sin Tendencia

tendencia no necesariamente obtenemos una serie estacionaria. Vale la pena decir que en ocasiones en la práctica puede ser necesario mas de una diferenciación ordinaria, es decir , aplicar d diferencias ordinarias a la serie $\{Y_t\}$ se define como:

$$\nabla^d Y_t = (1 - B)^d Y_t.$$

Note que la diferencia ordinaria de orden d es un filtro lineal invariante. Con lo anterior también podemos observar que la diferenciación ordinaria de orden 1 elimina una tendencia lineal. Adicionalmente uno puede verificar que una diferenciación ordinaria de orden 2 puede eliminar una tendencia cuadrática y así sucesivamente.

En general, los procedimientos que permite estimar y extraer las componentes de tendencia y/o estacional se conoce como suavizamiento. Existen varios métodos de suavizamiento, entre estos tenemos los presentamos en la siguiente sección.

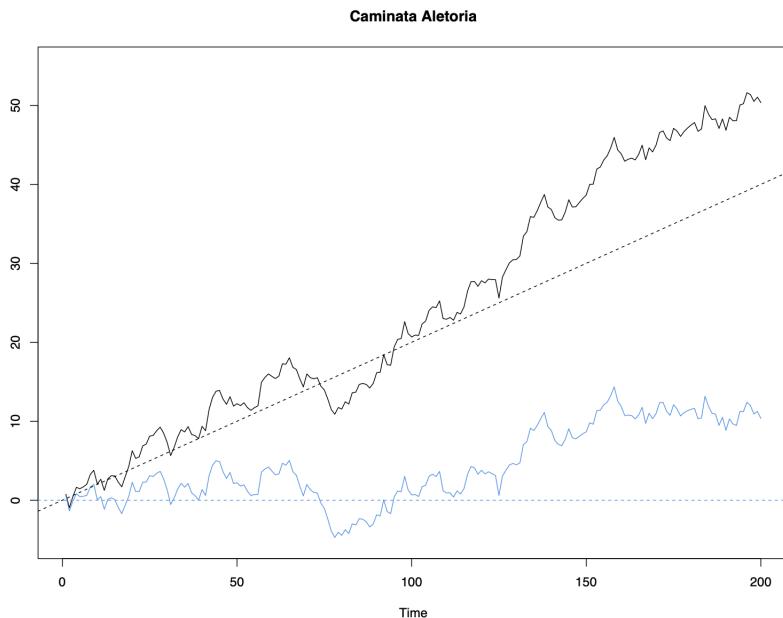


Figura 2.8: Caminata Aleatoria con y sin Drift

2.2.4 Suavizamiento de Series de Tiempo

Promedio Móvil

El promedio móvil es un método útil para descubrir ciertos rasgos en una serie de tiempo, como tendencias a largo plazo y componentes estacionales. En particular, si x_t representa las observaciones, entonces una forma de predecir predecir o estimar la tendencia de la serie es:

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

donde si $a_j = a_{-j} \geq 0$ y $\sum_{j=-k}^k a_j = 1$ se conoce como el promedio móvil simétrico de los datos. Este filtro puede extraer tanto la tendencia como la componente cíclica y eso dependerá del tipo de filtro usado. Un ejemplo de esto lo podemos ver para la serie **soi** 2.10.

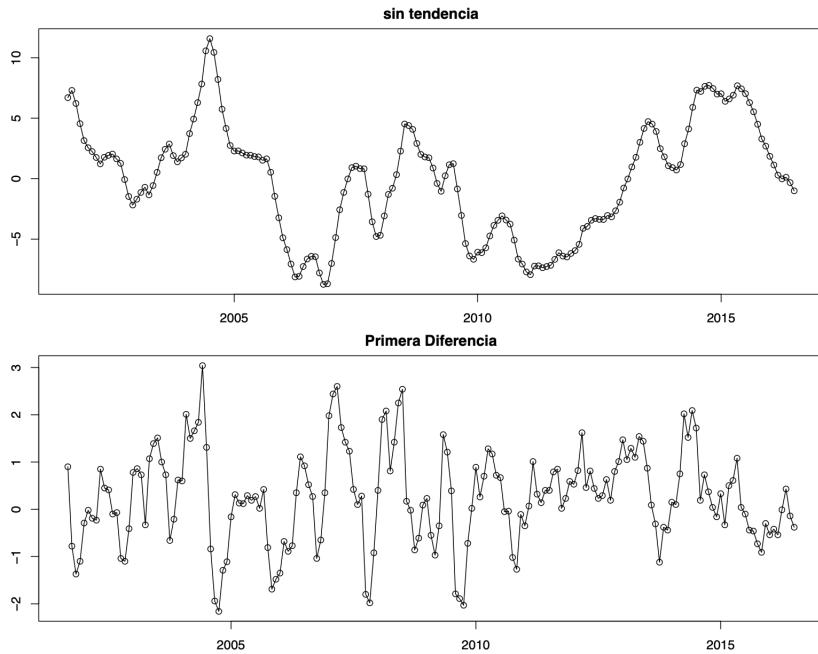


Figura 2.9: Series obtenidas después de corregir por la tendencia y luego de la diferenciación.

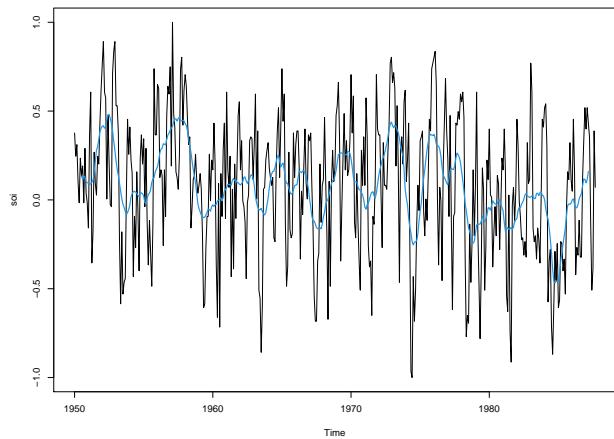


Figura 2.10: Ejemplo Suavizamiento usando promedio móviles para la serie soi

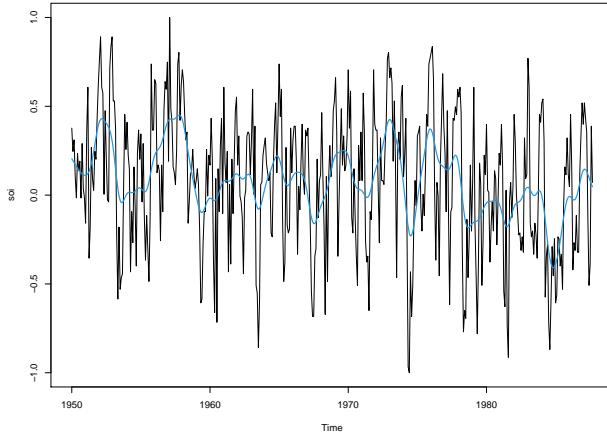


Figura 2.11: Suavizamiento Kernel soi

Suavizamiento Kernel

El suavizamiento kernel es un suavizador de promedio móvil que utiliza una función de ponderación, o kernel, para promediar las observaciones. Veamos ahora como queda el promedio móvil:

$$m_t = \sum_{i=1}^n w_i(t) x_i$$

donde

$$w_i(t) = K\left(\frac{t-i}{b}\right) / \sum_{j=1}^n K\left(\frac{t-j}{b}\right)$$

son los pesos y $K(\cdot)$ es una función kernel. Este estimador es llamado el Estimador de Nadaraya-Watson. Usaremos la función *ksmooth* de R, ver 2.11.

Loess o Lowess

Otro enfoque para suavizar un gráfico de tiempo es la regresión del vecino más cercano. La técnica se basa en la regresión de k vecinos más cercanos, en la que uno usa solo los datos $\{x_{t-k/2}, \dots, x_t, \dots, x_{t+k/2}\}$ para predecir x_t mediante regresión del tiempo, y luego establece $m_t = \hat{x}_t$. Primero, una cierta proporción de vecinos más cercanos a x_t para el tiempo t, se incluyen en un esquema de ponderación ($\nu_t = W\left(\frac{|t_i-t|}{\lambda_q(t)}\right)$); los valores más cercanos a

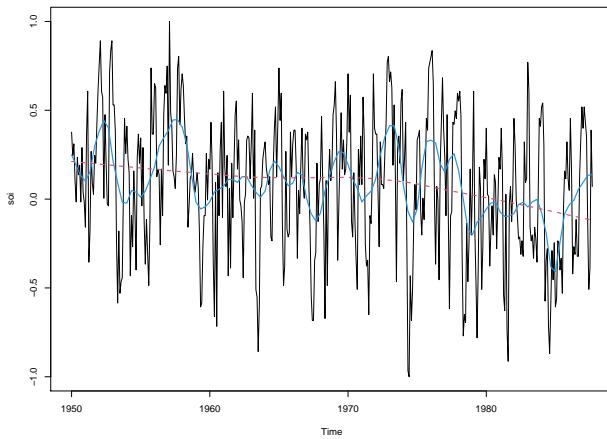


Figura 2.12: Ajuste Lowess a la serie soi

x_t en el tiempo obtienen más peso. Luego, se utiliza una regresión(polinomial de grado d , que es usualmente 1 o 2, es decir un ajuste localmente lineal o cuadrático) ponderada robusta(entre t y x_t) para predecir x_t y obtener los valores suavizados m_t . Cuanto mayor sea la fracción de vecinos más cercanos incluidos, más suave será el ajuste. El R se usa la función loess ver 2.12. La función $W(\cdot)$ es la función de peso tricúbica

$$W(u) = \begin{cases} (1 - u^3)^3, & \text{para } 0 \leq u < 1, \\ 0, & \text{para } u \geq 1. \end{cases}$$

y $\lambda_q(t)$ es la distancia de el q -ésimo tiempo mas lejano de t_i a t . Por supuesto que hay un procedimiento iterativo para hacer el ajuste si hay presencia de estacionalidad en los datos.

Suavizamiento Splines

Una forma obvia de suavizar los datos sería ajustar una regresión polinomial en términos del tiempo. Por ejemplo, un polinomio cúbico tendría $x_t = m_t + w_t$ donde $m_t = \beta_0 + \beta_1 t + \beta_2 t^2 + \beta_3 t^3$. Entonces podríamos ajustar m_t mediante mínimos cuadrados ordinarios.

Una extensión de la regresión polinomial es dividir primero el tiempo $t = 1, \dots, n$, en k intervalos, $[t_0 = 1, t_1], [t_1 + 1, t_2], \dots, [t_{k-1} + 1, t_k = n]$; los valores t_0, t_1, \dots, t_k se llaman nodos. Luego, en cada intervalo, se ajusta una

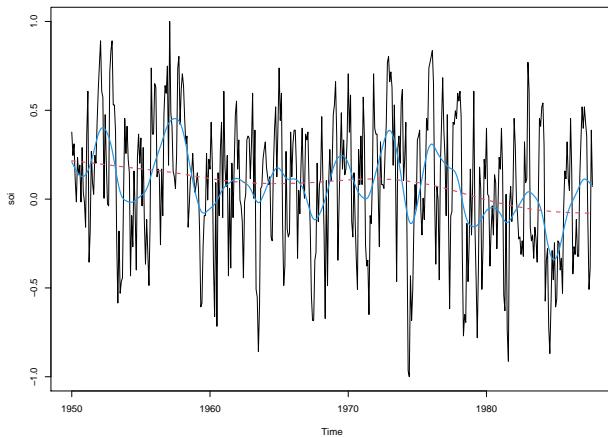


Figura 2.13: Splines Cúbicos soi

regresión polinomial, normalmente de orden 3, y esto se llama splines cúbicos. Un método relacionado es suavizar splines, que minimiza el compromiso entre el ajuste y el grado de suavidad dado por

$$\sum_{t=1}^n [x_t - m_t]^2 + \lambda \int (m_t'')^2 dt,$$

donde m_t es un spline cúbico con nodos en cada tiempo t y el grado de suavidad es controlado por $\lambda > 0$. El parámetro de suavizado en R es controlado por el argumento spar de la función smooth.spline del paquete stats, ver 2.13.

2.2.5 Transformación Box-Cox

En ocasiones, la serie de tiempo presenta una varianza marginal que cambia con respecto al tiempo, es decir, desde el punto de vista práctico, el rango de los valores que toma la variable de interés cambia con respecto al tiempo. Por lo general, ese cambio se presenta de forma monótona con respecto al tiempo. Cómo ejemplo veamos la serie del número de pasajeros en la gráfica 2.14, que ya fue descrita en el capítulo 1. En esta serie como lo pueden ustedes notar, el rango de valores se va haciendo cada vez mas grande. Es decir, la varianza marginal va creciendo con respecto al tiempo aumenta. En este caso, se siguiere hacer una transformación de potencia para estabilizar la varianza. Esta familia de transformaciones se llaman transformaciones

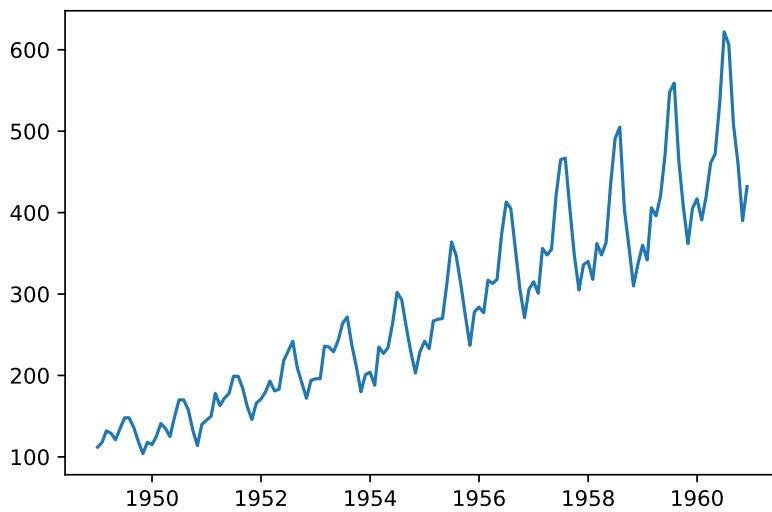


Figura 2.14: Serie del número de pasajero de una aerolínea medida mensualmente.

Box-Cox.

$$f_{\lambda}(u_t) = \begin{cases} \lambda^{-1}(u_t^{\lambda} - 1), & \text{si } u_t \geq 0, \text{ para } \lambda > 0, \\ \ln(u_t), & \text{si } u_t > 0, \text{ para } \lambda = 0. \end{cases} \quad (2.5)$$

Por supuesto, esta transformación tiene inversa y podemos en cualquier momento volver a la escala original de los datos. Otra forma de atacar la varianza no constante puede ser usando modelos no lineales que con llamados modelos de coeficientes que varían en el tiempo. Sin embargo, nosotros en éste curso trabajaremos la metodología basadas en transformaciones del tipo Box-Cox. La idea de la transformación consiste en elegir el valor de λ en 2.5 y aplicarlo a la serie que deseamos estabilizar la varianza. Python y R tienen rutinas que permiten darnos un idea de que valor podría tomar λ . Las rutinas están implementadas en las funciones `stats.boxcox` del módulo `scipy` en Python; y en `BoxCox.lambda` de la librería `forecast` o `BoxCox` de la librería `FitAR`.

Usualmente, si la serie presenta una varianza creciente o decreciente con respecto al tiempo, una transformación logarítmica estabiliza la varianza,

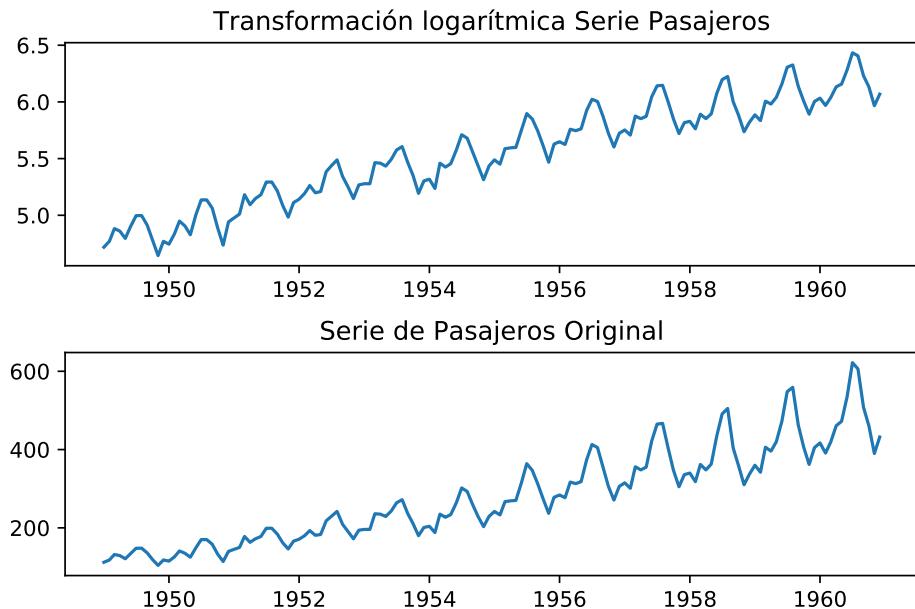


Figura 2.15: Serie Transformada y Serie Original

como es el caso de la serie de pasajeros, ver gráfica 2.15.

Note que si $\lambda = 1$ en 2.5 implica que no hay que hacer transformación Box-Cox.

Es importante señalar que existe una familia de transformaciones que permite trabajar con datos negativos y es debido a [43]. La transformación consiste de 2.16 :

$$\psi(\lambda, x) = \begin{cases} \{(x + 1)^\lambda - 1\}/\lambda & (x \geq 0, \lambda \neq 0), \\ \log(x + 1) & (x \geq 0, \lambda = 0), \\ -\{(-x + 1)^{2-\lambda} - 1\}/(2 - \lambda) & (x < 0, \lambda \neq 2), \\ -\log(-x + 1) & (x < 0, \lambda = 2). \end{cases}$$

Figura 2.16: Transformación YeoJhonson

Esta transformación está implementada en R(*VGAM::yeo.johnson* y *car::bcPower*) y Python(*sklearn.preprocessing.PowerTransformer* y *scipy.stats.yeojohnson*).

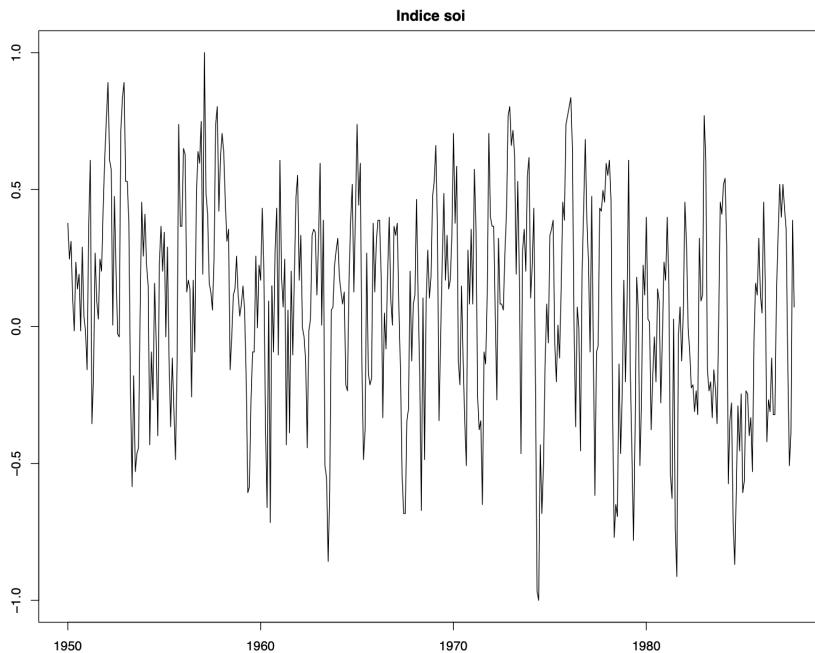


Figura 2.17: Gráficos índice soi

2.2.6 Diagramas de dispersión para la variable y sus retardos y el ACF muestral.

La idea de los diagramas de dispersión de la variable original y_t con sus propios retardos y_{t-k} es que podrían permitir la identificación de posibles relaciones no-lineales, inclusive podría ayudarnos a identificar posibles relaciones de y_t con retardos de otras variables. El ACF nos dice si existen sustanciales relaciones lineales entre la series y sus propios valores retardados, entonces se convierte también en una herramienta importante para explorar relaciones.

Un primer ejemplo de esto consiste en la exploración del índice soi s_t por sus siglas en inglés de *Southern Oscillation Index* y sus retardos. La gráfica 2.17 nos muestra la gráfica del índice soi mientras que la gráfica 2.18 nos muestra las gráficas de dispersión.

En el gráfico de dispersión podemos ver que se muestra un ajuste no paramétrico, de la posible relación entre las variables al igual que una estimación de la autocorrelación entre s_t y s_{t-h} . Vemos que varias de las relaciones exploradas parecen ser lineales. Uno puede observar que exis-

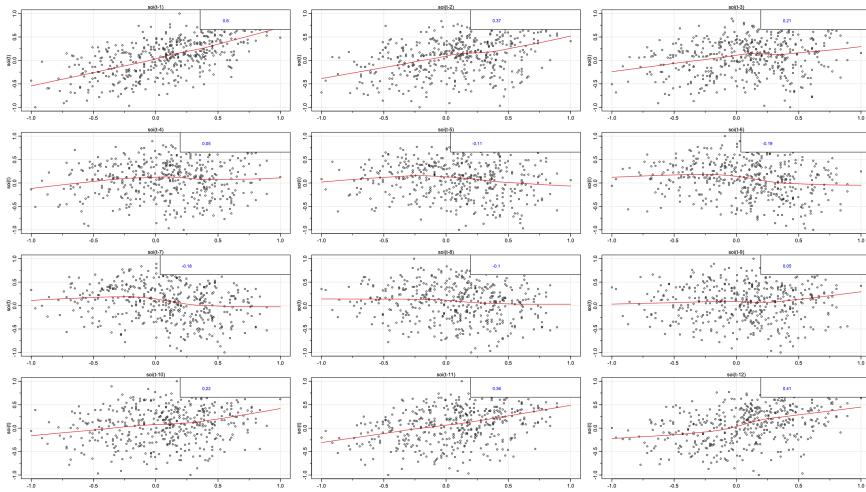


Figura 2.18: Gráfico de dispersión índice soi con sus propios retardos

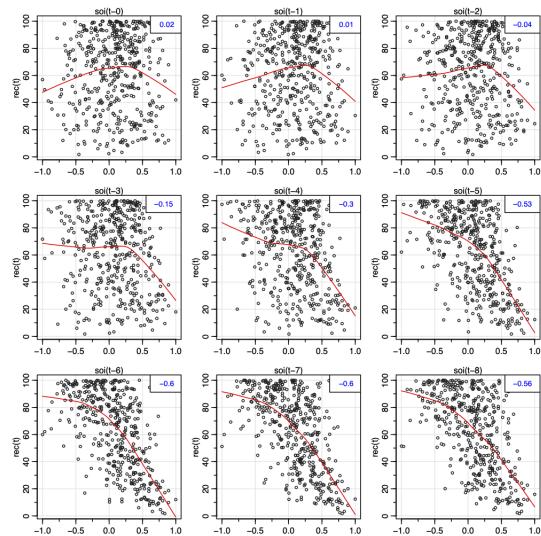


Figura 2.19: Gráficos de Dispersión Indice nuevos peces v.s retardos soi

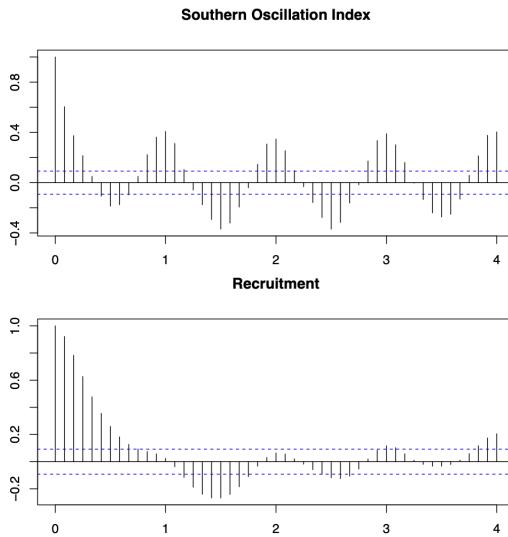


Figura 2.20: Autocorrelaciones para la serie soi y peces nuevos

ten relaciones lineales positivas en los rezagos $h = 1, 2, 11, 12$, mientras que negativas en los rezagos $h = 6, 7$, las demás parecen ser no significativas o no lineales. Note también que el gráfico de dispersión del índice de nuevos peces con retardos del indice soi 2.19 nos muestra tambien relaciones posiblemente lineales, al igual que no lineales. Ahora, podemos crear el gráfico de autocorrelación simple que nos permite estimar y graficar las autocorrelaciones para diferentes rezagos. En ambas gráficas 2.20 podemos ver que se muestran periodicidades en las correlaciones que corresponden a valores separados por 12 unidades. También podemos ver que las observaciones con 12 meses o un año de diferencia están fuertemente correlacionadas positivamente, al igual que las observaciones en múltiplos como 24, 36, 48, . . . Las observaciones separadas por seis meses están correlacionadas negativamente, lo que muestra que las excusiones positivas tienden a asociarse con las excusiones negativas a los seis meses ver [34] capítulo 1.

Note que la definición de autocorrelación para un único proceso estocástico puede ser extendida a dos o más procesos, la cual llamaremos autocorrelación cruzada. Veamos la definición para dos procesos. La función de autocovarianza cruzada para un proceso estocástico **conjuntamente estacionario**

cionario se define como

$$\gamma_{X,Y}(h) = Cov(X_{t+h}, Y_t) = E[(X_{t+h} - \mu_X)(Y_t - \mu_Y)].$$

La función de autocorrelación cruzada(CCF) para un proceso conjuntamente estacionario(X_t, Y_t) se define como

$$\rho_{X,Y}(h) = \frac{\gamma_{X,Y}(h)}{\sqrt{\gamma_X(0)\gamma_Y(0)}},$$

en donde se satisface que $\rho_{X,Y}(h) = \rho_{Y,X}(-h)$ porque $\gamma_{X,Y}(h) = \gamma_{Y,X}(-h)$.

Un estimador de la función de autocovarianza y autocorrelación cruzada es respectivamente

$$\hat{\gamma}_{X,Y}(h) = \frac{1}{n-h} \sum_{t=1}^n (x_{t+h} - \bar{x})(y_t - \bar{y}),$$

$$\hat{\rho}_{X,Y}(h) = \frac{\hat{\gamma}_{X,Y}(h)}{\sqrt{\hat{\gamma}_X(0)\hat{\gamma}_Y(0)}}.$$

La gráfica 2.21 nos muestra también periodicidades en la función de autocorrelación cruzada como en las autocorrelaciones simples. El pico alto es alcanzado en $h = -6$, este resultado implica que el SOI medido en el tiempo $t - 6$ meses está asociado con el reclutamiento serie en el tiempo t . Podríamos decir que SOI lidera la serie Recruitment por seis meses. El signo del CCF es negativo, lo que lleva a la conclusión de que las dos series se mueven en diferentes direcciones; es decir, los aumentos en SOI conducen a disminuciones en el reclutamiento y viceversa.

Información Mútua Promedio(AMI)

Ahora utilizaremos los paquetes de R nonlinearTseries y tseriesChaos para computar el average mutual information(AMI) o La información mutua promedio (AMI, la cual mide cuánto nos dice una variable aleatoria sobre otra, el cual se define como ver libro [18]:

$$I(X; Y) = \sum_i \sum_j p(x_i, y_j) \log_2 \left(\frac{p(x_i, y_j)}{p(x_i)p(y_j)} \right).$$

En el contexto del análisis de series de tiempo, AMI ayuda a cuantificar la cantidad de conocimiento obtenido sobre el valor de X_{t+d} al observar X_t . Equivalentemente, el AMI es una medida de qué tanto el conocimiento de X reduce la incertidumbre acerca de Y . Esto implica que $I(X, Y) = 0$ si y

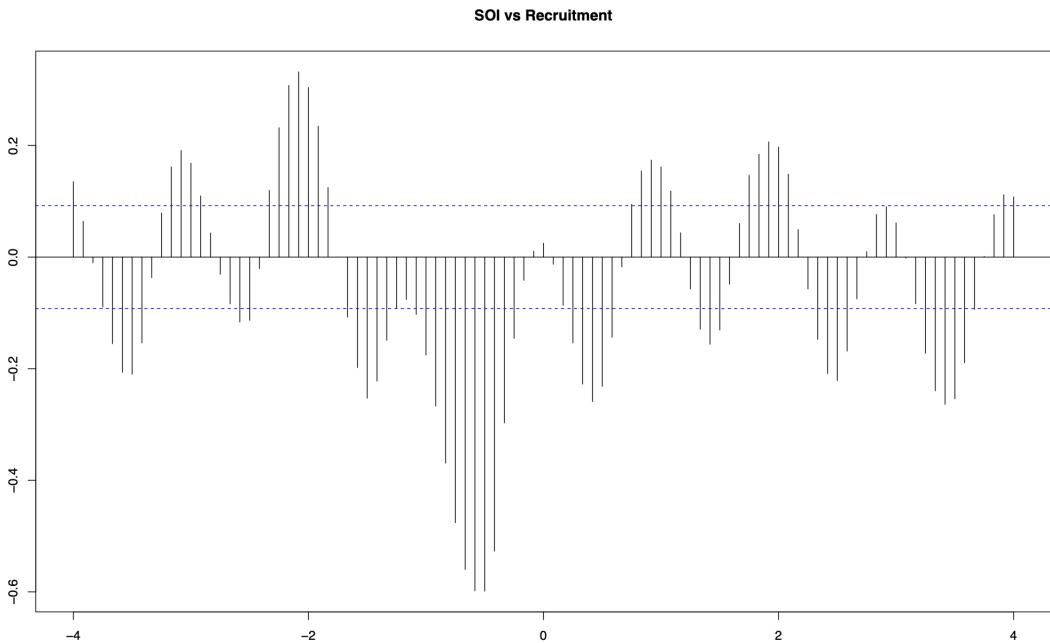


Figura 2.21: Función de autocorrelación cruzada soi v.s peces nuevos

sólo si X y Y son variables aleatorias independientes. $I(X; Y)$ describe la información que la medición X_t en el tiempo t aporta a la medición X_{t+d} en el tiempo $t+d$. Si se elige d como el valor alrededor del primer mínimo del AMI, entonces Y_t e y_{t+d} son parcialmente pero no totalmente independientes.

Nota 2.37. Existe mas formas de detectar la no-linealidad de una serie de tiempo estacionaria, por ejemplo:

- ✓ Usando la función de autocorrelación parcial(facp).
- ✓ Se puede usar la función de autocorrelación de distancia y cuatilograma como posibles medidas de la no linealidad.

2.2.7 Detección de Ciclos y Estacionalidades

Los componentes estacional y del ciclo describen eventos cíclicos a lo largo del tiempo, donde la duración de su ciclo los distingue. Como se mencionó anteriormente, el componente estacional tiene un ciclo constante, que se deriva y se vincula a la frecuencia en que se observa la serie. En este caso se

puede asumir que el modelo para mi serie de tiempos es como sigue:

$$X_t = \mu_t + S_t^{(s)} + a_t,$$

donde μ_t es la componente de tendencia y $S_t^{(s)}$ es la componente estacional.

En forma general la *estacionalidad* se representa formalmente como que el proceso $\{X_t\}$ satisface

$$E[X_t] = E[X_{t+s}]$$

para cada tiempo t y un entero s positivo llamado periodo. Éste capítulo se desarrollará siguiendo las pautas del libro [29]. Note que una serie que presente estacionalidad es no estacionaria. El periodo estacional, s , se define como número de observaciones que forman el ciclo estacional. Por ahora, suponemos que hay sólo existe un sólo tipo de estacionalidad. Porque por ejemplo en datos diarios podría existir una estacionalidad semana, con $s = 7$, otra mensual con $s = 30$, y otra anual, con $s = 365$.

Hay varias formas de introducir la estacionalidad en una serie de tiempo (estocástica y no estocástica). La forma mas simple que puede es a través de un efecto constante que se suma a los valores de la serie. Supongamos por ejemplo que la serie es una suma de una componente estacional $S_t^{(s)}$, y un proceso estacionario, n_t , así que el modelo para la serie observable Z_t será:

$$Z_t = S_t^{(s)} + n_t. \quad (2.6)$$

El proceso 6.3 es no estacionario ya que

$$E[Z_t] = E[S_t^{(s)}] + \mu,$$

donde μ es la media del proceso n_t . Note que como la componente estacional no toma el mismo valor en todos los periodos por definición 2.7. Ver ejemplo Serie de número de pasajeros, Ozono y lluvia.

Se puede considerar distintas hipótesis respecto al comportamiento del proceso estacional $S_t^{(s)}$. Por ejemplo que $S_t^{(s)}$ sea un proceso determinístico o determinista, es decir, una función constante para el mismo mes en distintos años:

$$S_t^{(s)} = S_{t+ks}^{(s)}, k = \pm 1, \pm 2, \dots,$$

como por ejemplo, los coeficientes estacionales pueden seguir una función sinusoidal. Estas funciones serán poco eficaces cuando la estacionalidad siga

un patrón determinista pero no sinusoidal, por ejemplo en series de producción, se muestran valores constantes todos los meses excepto en los meses de vacaciones, digamos diciembre y enero, así que el valor esperado es más bajo en esos meses. La idea en estos casos es introducir $11(s - 1)$ variables impulso o dummy $I_t^{(j)}$, que tomen 1 en un mes cero en el resto, es decir:

$$Z_t = \mu + \sum_{j=1}^{s-1} w_j I_t^{(j)} + n_t.$$

Otras series pueden no presentar un componente determinista, sino que la componente estacional evoluciona o tiene dinámica propia con respecto al tiempo. Lo cual supone que sigue un proceso estocástico, es decir, los factores estacionales no son constantes, pero siguen un proceso estacionario, oscilando alrededor de un valor medio de acuerdo a:

$$S_t^{(s)} = \mu^{(s)} + \eta_t,$$

donde $\mu^{(s)}$ es una constante que depende del mes y representa el efecto determinista de la estacionalidad y η_t es un proceso estocástico estacionario de media cero que introduce la variabilidad de cada año.

La otra forma de representar la estacionalidad consiste en que esta sea cambiante en el tiempo sin ningún valor medio fijo, es decir la estacionalidad sigue un proceso estocástico no estacionario, el cual se puede asumir que evoluciona como una caminata aleatoria:

$$S_t^{(s)} = S_{t-s}^{(s)} + \eta_t.$$

Por supuesto, pueden haber más tipos de estacionalidad, las cuales pueden seguir cualquier proceso estocástico no estacionario.

Por otro lado, la duración del ciclo del componente del ciclo no es necesariamente constante y típicamente puede variar de un ciclo al siguiente ver libro [24] capítulo 5. Existen varias herramientas para detectar componentes cíclicas y estacionales al igual que para su estimación o predicción.

Un Mapa de Calor podría ayudarnos en la detección de componentes estacionales y cíclicas. Por ejemplo, en el caso de la serie de gas natural, tenemos el siguiente mapa de calor 2.22:

En este ejemplo, el flujo de color de la Tasa de Desempleo es vertical, lo que indica el estado del ciclo. En este caso, las franjas verticales más claras representan el final de un ciclo y el comienzo del siguiente. Asimismo, las franjas verticales más oscuras representan los picos del ciclo.

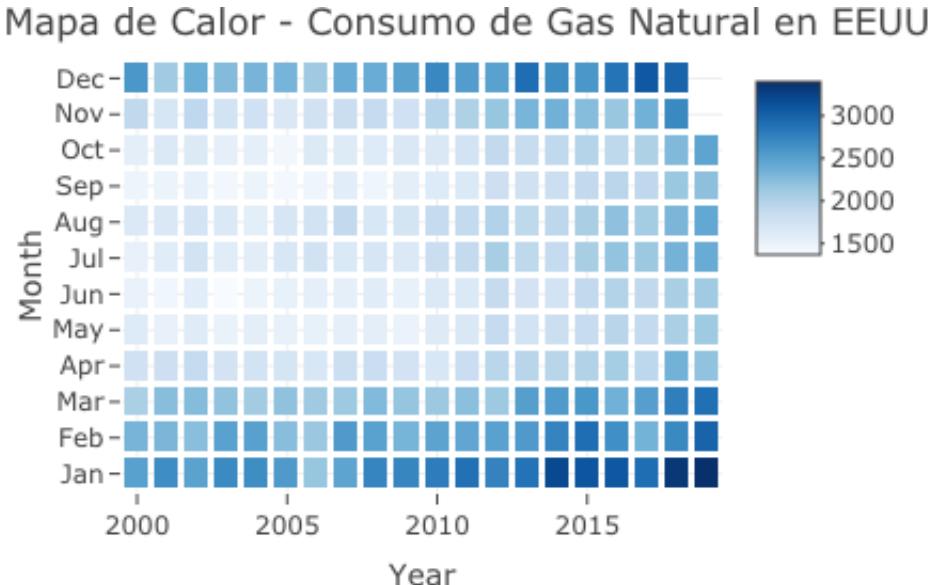


Figura 2.22: Mapa de Calor Serie Gas Natural

El Periodograma

Esta sección se desarrolla siguiendo el libro [29]. Del estudio de la estacionalidad de una serie de tiempo, pudimos ver que ésta componente es cíclica, la cual puede representarse por ejemplo mediante funciones periódicas. Note que Fourier demostró que una función $f(x)$ bajo ciertas condiciones puede escribirse como

$$f(x) = \frac{1}{2}a_0 + \sum_{n=1}^{\infty} a_n \cos(nx) + \sum_{n=1}^{\infty} b_n \sin(nx).$$

Por supuesto una función cíclica también puede expresarse mediante la anterior expansión. En la práctica un truncamiento es requerido para el modelamiento, es decir

$$f(x) \approx \frac{1}{2}a_0 + \sum_{n=1}^{N} a_n \cos(nx) + \sum_{n=1}^{N} b_n \sin(nx).$$

Note que el argumento de la función puede ser el tiempo, con lo cual se puede enmarcar en el contexto del análisis de series de tiempo. En seguida se considerará el caso para $N = 1$, pero en general puede extenderse a otros valores. Leer sección 7.4(Series de Fourier) del libro [19] para la implementación.

Supongamos que tenemos una serie de tiempo $\{Z_t\}$ la cual presenta una componente estacional y su modelo viene dado por:

$$Z_t = \mu + R \sin(\omega t + \theta) + a_t, \quad (2.7)$$

donde ω es la frecuencia angular dada por $\omega = \frac{2\pi}{s}$ donde s es el periodo estacional y R es la amplitud. La ecuación anterior puede re escribirse como

$$Z_t = \mu + R[\sin(\omega t) \cos \theta + \cos(\omega t) \sin \theta] + a_t,$$

donde por facilidad $\{a_t\} \sim RB(0, \sigma^2)$. Note que éste modelo puede verse como una regresión

$$Z_t = \mu + A \sin(\omega t) + B \cos(\omega t) + a_t \quad (2.8)$$

donde $A = R \cos \theta$ y $B = R \sin \theta$. Usted puede verificar que los estimadores de mínimos cuadrados de μ , A y B están dados por:

$\hat{\mu} = \frac{1}{T} \sum_{t=1}^T Z_t$, $\hat{A} = \frac{2}{T} \sum_{t=1}^T Z_t \sin(\omega t)$, $\hat{B} = \frac{2}{T} \sum_{t=1}^T Z_t \cos(\omega t)$ y así la amplitud estimada $\hat{R} = \hat{A}^2 + \hat{B}^2$, cual es cierto asumiendo que el periodo muestral T es un múltiplo del periodo s y por lo tanto $\sum_{t=1}^T \sin(\omega t) = 0$, $\sum_{t=1}^T \cos(\omega t) = 0$ y $\sum_{t=1}^T \sin(\omega t) \cos(\omega t) = 0$. Así, los residuos del modelo se calculan

$$\hat{a}_t = Z_t - \hat{\mu} - \hat{A} \sin(\omega t) + \hat{B} \cos(\omega t),$$

con lo cual, la varianza del ruido es

$$\hat{\sigma}^2 = \frac{1}{T} \sum_{t=1}^T \hat{a}_t^2.$$

Se puede verificar que la varianza estimada del proceso $\{Z_t\}$ está dada por

$$\frac{1}{T} \sum_{t=1}^T (Z_t - \mu)^2 = \frac{\hat{A}^2}{2} + \frac{\hat{B}^2}{2} + \hat{\sigma}^2 = \underbrace{\frac{\hat{R}^2}{2}}_{\text{Varianza debida a la onda}} + \hat{\sigma}^2$$

la cual es la descomposición de varianza en sus componentes ortogonales de variabilidad.

Exploración de Múltiples Círculos

La representación 2.8 de la sección anterior es adecuada cuando la estacionalidad es sinusoidal de periodo s , pero no sirve para describir funciones

periódicas generales. Fourier mostró que toda función periódica puede representarse como una suma de funciones sinusoidales de distinta amplitud y frecuencia. La idea entonces es generalizar el análisis anterior para un ciclo a suma de funciones armónicas con distintas frecuencias.

Dada la serie de longitud T , se denomina periodo básico o de Fourier, a las fracciones exactas del tamaño muestral, es decir

$$s_j = \frac{T}{j}, \quad j = 1, 2, \dots, T/2.$$

El valor máximo de periodo básico se obtiene para $j = 1$ y es T , es decir, observamos la onda una sola vez. El mínimo se obtiene para $j = T/2$ y es 2 ya que no se puede observar periodos que duren menos de 2 observaciones. Como antes, suele usarse las frecuencias en lugar de los periodos para el ajuste de los ciclos, así que las frecuencias básicas o de Fourier se define como

$$f_j = \frac{j}{T} \quad j = 1, 2, \dots, T/2.$$

Así, $1/T \leq f_j \leq 1/2$, y el valor máximo sería $f_j = 0.5$ y se conoce como la frecuencia Nyquist. Con esto, una serie de tiempo Z_t que es periódica se puede representar mediante

$$Z_t = \mu + \sum_{j=1}^{T/2} A_j \sin(\omega_j t) + \sum_{j=1}^{T/2} B_j \cos(\omega_j t) + a_t.$$

Note que éste modelo tiene tantos parámetros como observaciones, con lo cual hay que buscar un procedimiento para seleccionar las frecuencias que deben ser incluidas para explicar la evolución de la serie. Para esto se utilizará la herramienta conocida como el periodograma. Se puede verificar que la contribución de una onda a la varianza es dada por $\frac{R^2}{2}$, así que ondas asociadas con amplitudes grandes serán importantes en la explicación de la variabilidad de la serie, mientras que ondas asociadas con amplitudes bajas serán poco importantes. De forma análoga a como se mostró en la sección anterior, los estimadores de los coeficientes A_j y B_j están dados por

$$\hat{A}_j = \frac{2}{T} \sum_{t=1}^T Z_t \sin(\omega_j t)$$

$$\hat{B}_j = \frac{2}{T} \sum_{t=1}^T Z_t \cos(\omega_j t)$$

y

$$\hat{R}_j = \hat{A}_j^2 + \hat{B}_j^2,$$

donde $\omega_j = 2\pi f_j$ y $f = \frac{j}{T}$, con lo cual

$$TS_z^2 = \frac{T}{2} \sum_{j=1}^{T/2} \hat{R}_j^2.$$

Se le da el nombre de periodograma a la representación de cada $\frac{T\hat{R}_j^2}{2}$ en función de la frecuencia ω_j o f_j . Así

$$I(f_j) = \frac{T\hat{R}_j^2}{2}, \quad \text{con } 1/T \leq f_j \leq 1/2 \quad (2.9)$$

y

$$\bar{I} = \sum_{j=1}^{T/2} \hat{R}_j^2 = S_z^2.$$

Nos enfocaremos en las frecuencias básicas únicamente, lo cual no es restrictivo si T es muy grande, puesto que el número de frecuencias básicas será muy grande y siempre existirá alguna frecuencia básica muy próxima a la que puede interesarnos. El periodograma puede verse como una herramienta para la detección de posibles ciclos(ocultos) deterministas en una serie temporal. Por ejemplo, en una serie mensual estacional de periodo $s = 12$, esperamos encontrar un valor alto de periodograma para $f = 1/12$, pero también para $f = j/12$, es decir, $1/6, 1/4, 1/3$ que son armónicos del periodo estacional. Por otro lado, la serie puede tener otros ciclos no necesariamente ligado al periodo estacional, siendo el periodograma una buena herramienta para detectar estos posibles componentes.

Podemos considerar la amplitud calculada para la frecuencia f_j como un promedio de las amplitudes existentes en las frecuencias situadas en el intervalo $f_j \pm \frac{1}{2}T$. Con esto, se puede obtener el periodograma suavizado construyendo rectángulos con centro f_j , base igual a $1/T$ y alturas $I(f_j) = T\hat{R}_j^2/2$. Otra forma de suavizar el periodograma es

$$I(f) = \sum_{f_i-q}^{f_i+q} p_i I(f_i) \quad 0 \leq f \leq 0.5,$$

donde q representa la ventana usada y los p_i son los pesos simétricos y no modifica el valor de la varianza S_z^2 que es área bajo la curva. En el Archivo *Espectral.R* se encuentra algunos ejemplos

3

Procesos ARMA

Los procesos ARMA de su nombre en inglés *Autoregressive-Moving Average*, fueron popularizados por Box y Jenkins en su trabajo del año 1970, y han sido utilizados en múltiples campos.

Definición 3.1. Un proceso estocástico $\{X_t\}$ es ARMA(p, q) si es estacionario y es la solución de la ecuación en diferencias estocástica

$$X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \cdots + \phi_p X_{t-p} + Z_t + \theta_1 Z_{t-1} + \cdots + \theta_q Z_{t-q} \quad (3.1)$$

con $\phi_p \neq 0$ y $\theta_q \neq 0$, y $\{Z_t\} \sim RB(0, \sigma_z^2)$. Los parámetros p, q son los órdenes autoregresivos y de promedios móviles respectivamente.

Nota 3.2. La ecuación 3.1 se puede escribir en forma compacta introduciendo el operador de retardo B . El operador B se define como sigue:

$$BX_t = X_{t-1}$$

$$B^2 X_t = X_{t-2}$$

$$B^3 X_t = X_{t-3}$$

⋮

$$B^k X_t = X_{t-k}$$

y por definición $B^0 X_t = 1 \cdot X_t = X_t$, es decir, es el operador identidad. De esta manera, entonces la ecuación 3.1 se puede escribir como:

$$X_t - \phi_1 X_{t-1} - \phi_2 X_{t-2} - \cdots - \phi_p X_{t-p} = Z_t + \theta_1 Z_{t-1} + \cdots + \theta_q Z_{t-q}$$

y así, con base en el operador de retardo tenemos:

$$1 \cdot X_t - \phi_1 BX_t - \cdots - \phi_p B^p X_t = 1 \cdot Z_t + \theta_1 BZ_t + \cdots + \theta_q B^q Z_t$$

factorizando, tenemos

$$(1 - \phi_1 B - \cdots - \phi_p B^p)X_t = (1 + \theta_1 B + \cdots + \theta_q B^q)Z_t.$$

Así, si definimos los polinomios $\phi(B) = 1 - \phi_1 B - \cdots - \phi_p B^p$ y $\theta(B) = 1 + \theta_1 B + \cdots + \theta_q B^q$, entonces

$$\phi(B)X_t = \theta(B)Z_t.$$

A $\phi(B)$ se le conoce como el polinomio autoregresivo y a $\theta(B)$ el polinomio de promedios móviles.

Nota 3.3. Los procesos ARMA son una familia de procesos que sirven para modelar series de tiempo cuya función de autocorrelación o de autocovarianza muestral tienden a cero rápidamente.

Hay dos subfamilias de los modelos ARMA que son interesantes estudiarlas por separado, que son las familias de modelos autoregresivos puros y de promedios móviles puros.

Definición 3.4. Si en la definición 3.1 establecemos que $\phi(B) = 1$, entonces obtenemos

$$X_t = Z_t + \theta_1 Z_{t-1} + \cdots + \theta_q Z_{t-q} \quad (3.2)$$

el cual es llamado el proceso MA(q) o de promedios móviles de orden q .

Note que la ecuación en diferencias estocástica tiene solución trivial $\theta(B)Z_t$. Una solución de una ecuación en diferencias estocástica en un proceso estocástico. Para detalles de las ecuaciones en diferencias ver [11] página 41, [29] Página 137, [2] Página 105 y [31] Página 121.

Ejemplo 3.5. Como caso particular consideremos el proceso MA(1), es decir,

$$X_t = Z_t + \theta Z_{t-1}.$$

Note que el proceso MA(1) es estacionario en el sentido débil.

$$E[X_t] = E[Z_t + \theta Z_{t-1}] = E[Z_t] + \theta E[Z_{t-1}] = 0.$$

Ahora computemos la función de autocovarianza $Cov(X_{t+h}, X_t)$. Para $h = 0$, tenemos que

$Cov(X_t, X_t) = E[X_t^2] = E[(Z_t + \theta Z_{t-1})^2] = E[Z_t^2 + 2\theta Z_t Z_{t-1} + \theta^2 Z_{t-1}^2]$,
así, $Cov(X_t, X_t) = \sigma_z^2 + \theta^2 \sigma_z^2 = (1 + \theta^2) \sigma_z^2$.

Ahora para $h = 1$, $Cov(X_{t+1}, X_t) = E[(Z_{t+1} + \theta Z_t)(Z_t + \theta Z_{t-1})] = \sigma_z^2 \theta ..$

Note que para $h = 2$, tenemos que

$$Cov(X_{t+2}, X_t) = E[(Z_{t+2} + \theta Z_{t+1})(Z_t + \theta Z_{t-1})] = 0.$$

Note que para $h \geq 2$ se tiene $Cov(X_{t+h}, X_t) = 0$, entonces resumiendo tenemos que

$$\gamma(h) = \begin{cases} (1 + \theta^2) \sigma_z^2, & \text{si } h = 0, \\ \sigma_z^2 \theta, & \text{si } h = 1, \\ 0, & \text{si } h \geq 2. \end{cases}$$

Cómo queda la función de autocorrelación?

En general, para procesos $MA(q)$ tenemos también que es estacionario con $E[X_t] = 0$ y función de autocovarianza dada por

$$\gamma(h) = \begin{cases} \sigma_z^2 \sum_{j=0}^{q-|h|} \theta_j \theta_{j+|h|}, & \text{si } |h| \leq q, \\ 0, & \text{si } |h| > q. \end{cases}$$

Ejercicio 3.6. Hacer una simulaciones de los proceso $MA(1)$ y $MA(2)$ para diferentes tamaños de muestra $T = 100, 300, 500$ y computar la función de autocorrelación muestral para cada caso. Compare la función de autocorrelación muestral con la teórica.

Definición 3.7. Consideremos ahora que en la definición 3.1 establecemos que $\theta(B) = 1$, entonces tenemos

$$X_t = \phi_1 X_{t-1} + \cdots + \phi_p X_{t-p} + Z_t,$$

que se conoce como el proceso $AR(p)$ o autoregresivo puro de orden p .

Ejemplo 3.8. Como ejemplo consideremos de la definición anterior que $p = 1$, es decir, que establecemos el proceso $AR(1)$

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

Establezcamos que $|\phi| < 1$ y que Z_t es no correlacionado con X_s para cada $s < t$, entonces no es difícil verificar que

$$X_t = \sum_{j=0}^{\infty} \phi^j Z_{t-j}$$

es una solución de la ecuación en diferencias $X_t = \phi X_{t-1} + Z_t$.

A continuación se encuentra el bosquejo de las deducciones para el ejemplo 3.8.

En efecto, dado que la sucesión $\{\phi^j\}$ es absolutamente sumable $\sum_{j=0}^{\infty} |\phi^j| = \frac{1}{1-|\phi|}$, entonces $\sum_{j=0}^{\infty} \phi^j Z_{t-j}$ está bien definido, además es una solución de $X_t = \phi X_{t-1} + Z_t$ la cual es equivalente a $X_t - \phi X_{t-1} = Z_t$.

Note que al reemplazar en la ecuación anterior a X_t por $\sum_{j=0}^{\infty} \phi^j Z_{t-j}$ que es la solución propuesta, tenemos que

$$\begin{aligned} & \sum_{j=0}^{\infty} \phi^j Z_{t-j} - \phi \sum_{j=0}^{\infty} \phi^j Z_{t-1-j} \\ &= Z_t + \phi Z_{t-1} + \phi^2 Z_{t-2} + \cdots - \phi(Z_{t-1} + \phi Z_{t-2} + \phi^2 Z_{t-3} + \cdots) \\ &= Z_t + \cancel{\phi Z_{t-1}} + \cancel{\phi^2 Z_{t-2}} + \cdots - \cancel{\phi Z_{t-1}} - \cancel{\phi^2 Z_{t-2}} - \phi^3 Z_{t-3} - \cdots \\ &= Z_t \end{aligned}$$

Lo cual confirma que es la solución. Note que en [3], pág. 79, se puede ver una forma para encontrar la solución de la ecuación en diferencias $X_t = \phi X_{t-1} + Z_t$.

Note que $X_t = \sum_{j=0}^{\infty} \phi^j Z_{t-j}$ es estacionario de media cero y facv

$$\gamma_X(h) = \sigma_Z^2 \sum_{j=0}^{\infty} \phi^j \phi^{j+h} = \sigma_Z^2 \phi^h \sum_{j=0}^{\infty} (\phi^2)^j = \frac{\sigma^2 \phi^h}{1 - \phi^2}$$

Usando la ecuación 2.3 se puede verificar que $\sum_{j=0}^{\infty} \phi^j Z_{t-j}$ es la única solución estacionaria de $X_t = \phi X_{t-1} + Z_t$.

Para esto, sea $\{Y_t\}$ otra solución estacionaria de $Y_t = \phi Y_{t-1} + Z_t$. Entonces tenemos lo siguiente:

$$\begin{aligned} Y_t &= \phi Y_{t-1} + Z_t && \text{iterando, tenemos} \\ &= \phi(\phi Y_{t-2} + Z_{t-1}) + Z_t \\ it.1 &= Z_t + \phi Z_{t-1} + \phi^2 Y_{t-2} \\ &\vdots \\ it.k &= Z_t + \phi Z_{t-1} + \cdots + \phi^k Z_{t-k} + \phi^k Y_{t-k-1} \end{aligned}$$

Como $\{Y_t\}$ es estacionario, entonces $E[Y_t^2]$ es finito y no depende de t , así que

$$Y_t - \sum_{j=0}^k \phi^j Z_{t-j} = \phi^k Y_{t-k-1}$$

elevando al cuadrado y tomando el valor esperado tenemos que

$$E \left[\left(Y_t - \sum_{j=0}^k \phi^j Z_{t-j} \right)^2 \right] = \phi^{2k} E[Y_{t-k-1}^2]$$

haciendo tender a infinito a k , tenemos que $\phi^{2k} \rightarrow 0$ y $E[Y_{t-k-1}^2] < \infty$. Así que

$$E \left[\left(Y_t - \lim_{k \rightarrow \infty} \sum_{j=0}^k \phi^j Z_{t-j} \right)^2 \right] = 0$$

entonces $\lim_{k \rightarrow \infty} \sum_{j=0}^k \phi^j Z_{t-j} = \sum_{j=0}^{\infty} \phi^j Z_{t-j}$ converge en M.C. a Y_t , que es el mismo X_t , así que el límite es el mismo.

¿Qué sucede si $|\phi| > 1$? Se puede verificar que $\sum_{j=0}^{\infty} \phi^j Z_{t-j}$ no converge con ningún criterio, pero si consideramos la ecuación $X_t = \phi X_{t-1} + Z_t$ de la forma $\phi X_{t-1} = X_t - Z_t$. Entonces $X_{t-1} = \phi^{-1} X_t - \phi^{-1} Z_t$, así $X_t = \phi^{-1} X_{t+1} - \phi^{-1} Z_{t+1}$. De nuevo, iterando hasta $k+1$ tenemos

$$X_t = -\phi^{-1} Z_{t+1} - \cdots - \phi^{-k-1} Z_{t+k+1} + \phi^{-k-1} X_{t+k+1}$$

Con los mismos argumentos que para $|\phi| < 1$, tenemos que

$$X_t = - \sum_{j=1}^{\infty} \phi^{-j} Z_{t+j} \quad \text{converge en M.C.}$$

Es la única solución estacionaria de $X_t = \phi X_{t-1} + Z_t$ para $|\phi| > 1$.

Nota 3.9. Si $|\phi| = \pm 1$ la ecuación

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

no tiene solución estacionaria. Cuando $\phi = 1$ el proceso tiene tendencia estocástica como la de una caminata aleatoria, lo cual también se entiende que tiene una raíz unitaria, es decir, la ecuación anterior se reduce a

$$(1 - B)X_t = Z_t,$$

con lo cual el polinomio autoregresivo $\phi(B) = 1 - B$, que puede verse como un polinomio en la indeterminada $z \in \mathbb{C}$, $\phi(z)$, tiene una raíz en $z = 1$.

Si $\phi = -1$, el proceso producido es explosivo.

Nota 3.10. Para el caso de modelos ARMA generales, el estudio de la estacionariedad recae sobre la función racional

$$\frac{\theta(z)}{\phi(z)},$$

en el sentido si esta se puede expandir como una serie de potencias en z . Esta expansión recae sobre raíces del polinomio autorregresivo $\theta(z)$ como se puede verificar en la siguiente sección.

Ejercicio 3.11. Hacer simulaciones para los diferentes escenarios, tanto en R como en Python.

Antes de continuar con mas resultados de los procesos ARMA , veamos unos resultados de expansiones de funciones racionales que van a ser necesarios en las siguientes secciones. Detalles puede ser encontrados del libro [31] capítulo 3.

3.1 Expansión de funciones Racionales

Como veremos en los capítulos subsecuentes, muchos de los modelos de series de tiempo contienen funciones de transferencia de la forma $Y_t = \frac{\delta(B)}{\gamma(B)}X_t$. Para que ésta expresión tenga sentido es necesario que el operador $\frac{\delta(B)}{\gamma(B)}$ cumpla una condición de estabilidad. Es decir, Y_t deberá ser “acotado” siempre que X_t se acotado, lo cual se traduce en que la expansión de la serie(de potencia) $\{\omega_0 + \omega_1 z + \omega_2 z^2 + \dots\}$ de $\omega(z) = \frac{\delta(z)}{\gamma(z)}$ es que sea convergente siempre que $|z| \leq 1$. Entonces, será necesario verificar si la serie es o no convergente. La convergencia o no de la serie puede ser verificada expresando el cociente $\frac{\delta(z)}{\gamma(z)}$ como suma de fracciones parciales.

Por ejemplo, consideremos que el denominador $\gamma(z)$ se puede factorizar como sigue

$$\gamma(z) = \gamma_m \prod (z - \lambda_i) = \gamma_0 \prod \left(1 - \frac{z}{\lambda_i}\right)$$

cuyas raíces pueden ser todas complejas. Entonces asumiendo que no hay raíces repetidas, y tomando $\gamma_0 = 1$ sin pérdida de generalidad, el cociente $\frac{\delta(z)}{\gamma(z)}$ puede ser escrito como suma de fracciones parciales

$$\frac{\delta(z)}{\gamma(z)} = \frac{\kappa_1}{1 - z/\lambda_1} + \frac{\kappa_2}{1 - z/\lambda_2} + \dots + \frac{\kappa_m}{1 - z/\lambda_m}.$$

Así, la función racional converge si y sólo si, la expansión cada una de sus sumas parciales en término de potencias ascendentes de z (“series de po-

tencia"). Para que la expansión

$$\frac{\kappa}{1-z/\lambda} = \kappa \left\{ 1 + \frac{z}{\lambda} + \left(\frac{z}{\lambda}\right)^2 + \left(\frac{z}{\lambda}\right)^3 + \dots \right\}$$

sea convergente para todo $|z| \leq 1$, es necesario y suficiente que $|\lambda| > 1$, es decir, λ es la raíz del polinomio $1 - z/\lambda$.

En general lo que se puede establecer es lo siguiente:

Proposición 3.12. *La expansión $\omega(z) = \omega_0 + \omega_1 z + \omega_2 z^2 + \dots$ de la función racional $\delta(z)/\gamma(z)$ converge para todo $|z| \leq 1$ si y sólo si cada raíz(polo, si son en general funciones de la variable compleja) λ de $\gamma(z) = 0$ cae por fuera del círculo unitario tal que $|\lambda| > 1$, es decir si están por fuera del círculo unitario complejo.*

Nota 3.13. *La expansión de la función racional puede converger bajo condiciones las cuales son mas o menos exigentes, en las restricciones que se imponen sobre $|z|$. En efecto, para cualquier serie*

$$\omega(z) = \omega_0 + \omega_1 z + \omega_2 z^2 + \dots,$$

entonces existe un número real $r \geq 0$, llamado el radio de convergencia, tal que si $|z| < r$, entonces la serie converge absolutamente con $\sum |\omega_i| < \infty$, mientras que si $|z| > r$, a serie diverge.

En el caso de la función racional $\delta(z)/\gamma(z)$, la condición de convergencia de la expansión es que $|z| \leq r = \min\{|\lambda_1|, |\lambda_2|, \dots, |\lambda_m|\}$, donde λ_i son todas las raíces de $\gamma(z) = 0$.

A veces la condición de convergencia de la expansión es dada en términos de la función $\delta(z^{-1})/\gamma(z^{-1})$, y es que $|z| > r = \max\{|\mu_1|, \dots, |\mu_m|\}$ donde $\mu_i = 1/\lambda_i$ es una raíz de $\gamma(z^{-1}) = 0$. Lo anterior hace que los resultados dados para la expansión en términos de potencias negativas de z se inviertan, es decir, que las raíces de $\gamma(z^{-1})$ deben estar por dentro del círculo unitario.

Usando la función `UnitCircle::uc.check` de R, podemos obtener una salida para chequear si las raíces de un polinomio está por fuera del círculo unitario complejo. Un ejemplo está dado en la gráfica 3.1.

3.1.1 Relaciones de Recurrencia

La sección anterior nos establece que si z no es una raíz de $\alpha(z)$, entonces $\frac{\beta(z)}{\alpha(z)}$ tiene una expansión en potencias en z , donde $\alpha()$ y $\beta()$ son polinomios de grado p y k respectivamente, como sigue:

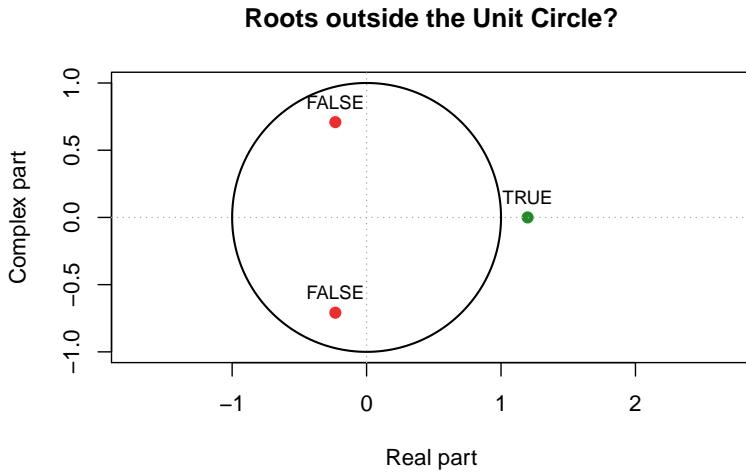


Figura 3.1: Gráfica de las raíces del polinomio $-3x^3 + 2.2x^2 + 2$

$$\frac{\beta(z)}{\alpha(z)} = \omega(z) = \{\omega_0 + \omega_1 z + \omega_2 z^2 + \dots\}.$$

Pero los valores de los coeficientes de la expansión son desconocidos. Sin embargo, se puede obtener una relación de recurrencia debido a la igualdad de los polinomios infinitos convergentes. Es decir, si tenemos la relación

$$\frac{\beta(z)}{\alpha(z)} = \{\omega_0 + \omega_1 z + \omega_2 z^2 + \dots\},$$

esta es equivalente a

$$\beta(z) = \alpha(z)\{\omega_0 + \omega_1 z + \omega_2 z^2 + \dots\},$$

la cual haciendo el producto al lado derecho, y utilizando la igualdad de polinomios, tenemos las siguientes relaciones de recurrencia:

$$\begin{aligned}\beta_j &= \sum_{i=0}^r \alpha_i \omega_{j-i}; \quad 0 \leq j \leq k, \\ 0 &= \sum_{i=0}^r \alpha_i \omega_{j-i}; \quad j > k,\end{aligned}\tag{3.3}$$

donde $r = \min(p, j)$. Las últimas ecuaciones se pueden resolver para ω_j , lo

cual se obtiene que los coeficientes de las expansión son:

$$\begin{aligned}\omega_j &= (\beta_j - \sum_{i=1}^r \alpha_i \omega_{j-i}) / \alpha_0; \quad 0 \leq j \leq k, \\ \omega_j &= - \sum_{i=1}^r \alpha_i \omega_{j-i} / \alpha_0; \quad j > k.\end{aligned}\tag{3.4}$$

Las primeras $k + 1$ recurrencias son las condiciones iniciales, de ahí en adelante, todos los valores de ω_j se obtiene de forma recurrente.

Nota 3.14. Leer la sección de Series de Laurent del libro [31], que empieza en la página 67.

En su forma mas general, uno pueden considerar una expansión en series de Taylor o Laurent, de funciones analíticas de la variable compleja.

3.2 El modelo ARMA(1,1)

Antes de entrar con unos resultados mas generales para los modelos $ARMA(p, q)$, nos enfocaremos en un primer modelo introductorio que es el modelo ARMA(1,1).

Sea $\{X_t\}$ un proceso estocástico que sigue un modelo $ARMA(1,1)$, es decir, es solución estacionaria de la ecuación en diferencias estocástica

$$X_t - \phi X_{t-1} = Z_t + \theta Z_{t-1},$$

donde $\{Z_t\} \sim RB(0, \sigma^2)$ y $\phi + \theta \neq 0$.

En términos del operador de retardo B, el proceso $ARMA(1,1)$ puede escribirse como

$$\phi(B)X_t = \theta(B)Z_t,$$

donde $\phi(B) = 1 - \phi B$ y $\theta(B) = 1 + \theta B$.

Veamos bajo que condiciones podemos obtener una solución estacionaria. Si $|\phi| < 1$ entonces la función racional $\theta(z)/\phi(z)$ tiene expansión en potencias crecientes para z , es decir:

$$\frac{\theta(z)}{\phi(z)} = \omega(z) = (\omega_0 + \omega_1 z + \omega_2 z^2 + \dots)$$

cuya recurrencia puede ser encontrada usando 3.4, ya que la raíz de $\phi(z) = 1 - \phi z = 0$, que es $z = 1/\phi$; su valor absoluto $|z| = |1/\phi| > 1$, luego está por fuera del círculo unitario. Entonces:

$$\omega_0 = 1, \quad \text{ya que } r = \min(1, 0) = 0,$$

es decir, $\omega_0 = 1$.

$$\omega_1 = \theta - (-\phi)\omega_0 = \theta + \phi, \quad \text{ya que } r = \min(1, 1) = 1.$$

Note ahora que

$$\omega_2 = -(-\phi\omega_1) = \phi(\theta + \phi), \quad \text{ya que } r = \min(1, 2) = 1.$$

En general se puede ver que

$$\omega_j = \phi^{j-1}(\phi + \theta), \quad \text{y } \omega_0 = 1.$$

Note que la serie $\{\omega_j\}$ es absolutamente convergente, y así el operador $\frac{\theta(B)}{\phi(B)}$ está bien definido y se puede escribir como

$$\frac{\theta(B)}{\phi(B)} = \omega_0 + \omega_1 B + \omega_2 B^2 + \cdots = \omega(B).$$

Entonces, la serie

$$X_t = \frac{\theta(B)}{\phi(B)} Z_t = \omega(B) Z_t = Z_t + (\phi + \theta) \sum_{j=1}^{\infty} \phi^{j-1} Z_{t-j}. \quad (3.5)$$

En otras palabras, lo que hemos hecho es encontrar una solución de la ecuación en diferencias, y además esa solución es estacionaria.

Veamos ahora que sucede si $|\phi| > 1$. Note que la ecuación autoregresiva

$$X_t = \phi X_{t-1} + Z_t + \theta Z_{t-1}$$

puede escribirse como $X_{t+1} = \phi X_t + Z_{t+1} + \theta Z_t$, así tenemos que $\phi X_t = X_{t+1} - Z_t - \theta Z_{t-1}$, con lo cual $X_t = \frac{1}{\phi} X_{t+1} - \frac{1}{\phi} Z_t - \frac{1}{\phi} \theta Z_{t-1}$, con lo cual tenemos que la ecuación autoregresiva es equivalente a

$$X_t - \frac{1}{\phi} X_{t+1} = \left(-\frac{1}{\phi}\right)(Z_{t+1} + \theta Z_t),$$

la cual escrita en forma compacta es

$$(1 - \frac{1}{\phi} B^{-1}) X_t = \left(-\frac{1}{\phi}\right)(B^{-1} + \theta) Z_t$$

donde $B^{-j} X_t = X_{t+j}$. Así, el cociente de polinomios

$$\frac{z^{-1} + \theta}{1 - 1/\phi z^{-1}}$$

puede expandirse como una series de potencias negativas

$$\omega_0 + \omega_1 z^{-1} + \omega_2 z^{-2} + \dots = \frac{\beta(z)}{\alpha(z)},$$

siempre que las raíces de $\alpha(z)$ estén por dentro del círculo unitario. Para éste caso $\beta(z) = z^{-1} + \theta$ y $\alpha(z) = 1 - \frac{1}{\phi}z^{-1}$, así que la única raíz de $1 - \frac{1}{\phi}z^{-1} = 0$ es $z = \frac{1}{\phi}$, la cual está por dentro del círculo unitario dado que $|\phi| > 1$. Note que la raíz $1/\phi$ de $\alpha(z)$, es conocido como un polo de la función $\frac{\beta(z)}{\alpha(z)}$.

No es inmediato, pero se puede ver que

$$\left(-\frac{1}{\phi}\right) \frac{z^{-1} + \theta}{1 - 1/\phi z^{-1}} = -\theta\phi^{-1}Z_t - (\phi + \theta) \sum_{j=1}^{\infty} \phi^{-j-1}Z_{t+j}.$$

Así, la única solución estacionaria de $X_t = \phi X_{t-1} + Z_t + \theta Z_{t-1}$ con $|\phi| > 1$ se puede escribir como

$$X_t = -\theta\phi^{-1}Z_t - (\phi + \theta) \sum_{j=1}^{\infty} \phi^{-j-1}Z_{t+j}, \quad (3.6)$$

la cual depende de ruidos presentes y futuros.

Finalmente se puede verificar que si $\phi = \pm 1$, entonces no hay solución estacionaria de

$$X_t = \phi X_{t-1} + Z_t + \theta Z_{t-1},$$

por lo tanto no hay un proceso ARMA(1,1) con $\phi = \pm 1$ ya que debe ser estacionario.

En resumen:

- ✓ Existen soluciones estacionarias y no estacionarias de las ecuaciones del proceso ARMA(1,1) dependiendo del valor de ϕ .
- ✓ Si $|\phi| < 1$ la única solución estacionaria está dada en 3.5 y es llamada una *solución causable o función causable* de $\{Z_t\}$, ya que X_t puede ser expresado en términos de de valores pasados y presentes de $Z_s, s \leq t$.
- ✓ Si $|\phi| > 1$, la única solución estacionaria es 3.6, la cual es la solución no causable puesto que X_t es una función de $Z_s, s \geq t$.
- ✓ Si $\phi = \pm 1$, entonces no se tiene solución estacionaria. Ver ejercicio 2.8 del libro [3].

Nota 3.15. Note que se puede encontrar la función de autocovarianza teórica de proceso ARMA(1,1) usando 2.3.

En seguida daremos resultados generales de los procesos ARMA.

3.3 Características de los Procesos ARMA(p,q) Generales

Empezaremos esta sección hablando del concepto de causalidad de los procesos ARMA(p,q). La idea de causalidad es que el proceso X_t depende únicamente de variables del proceso de ruido $\{Z_t\}$ del pasado y del presente. Esto es porque tiene mas sentido desde el punto de vista práctico que se dependa de variables de ruido en el pasado y presente, que se dependa también del futuro de las variables de ruido.

Definición 3.16. *Un proceso ARMA(p, q) definido por las ecuaciones*

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

es llamado causable o función causable de $\{Z_t\}$ si existe una sucesión de constantes $\{\psi_j\}$ tal que $\sum |\psi_j| < \infty$ y

$$X_t = \sum_{j=0}^{\infty} \psi_j Z_{t-j},$$

para cada $t \in \mathbb{Z}$.

En seguida tenemos un resultado que nos permite comprobar de manera fácil si un proceso ARMA es o no causable.

Teorema 3.17. *Sea X_t un proceso ARMA(p, q) para el cual los polinomios $\phi(\cdot)$ y $\theta(\cdot)$ no tienen ceros en común. Entonces X_t es causable si y solo si $\phi(z) \neq 0$ para todo $z \in \mathbb{C}$ tal que $|z| \leq 1$, los coeficientes $\{\psi_j\}$ en $X_t = \sum_{j=0}^{\infty} \psi_j Z_{t-j}$ son determinados por la relación:*

$$\psi(z) = \sum_{j=0}^{\infty} \psi_j z^j = \frac{\theta(z)}{\phi(z)}, |z| \leq 1.$$

Este teorema lo dejamos sin demostración, pero la prueba subyace de las ideas de sección 2.3. Note que chequear la causalidad consiste únicamente en ver si las raíces del polinomio $\phi(z)$ están por fuera del círculo unitario complejo.

Ejercicio 3.18. *Cuál de los siguientes procesos son causables?*

◊ $X_t - 0.2X_{t-1} - 0.48X_{t-2} = Z_t$

$$\diamond X_t + 1.8X_{t-1} + 0.81X_{t-2} = Z_t + 0.8Z_{t-1}$$

Teorema 3.19. Si $\phi(z) \neq 0$ para todo $z \in \mathbb{C}$ tal que $|z| = 1$ entonces las ecuaciones ARMA

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

tiene la única solución estacionaria

$$X_t = \sum_{j=-\infty}^{\infty} \psi_j Z_{t-j}$$

donde los coeficientes ψ_j son determinados por

$$\frac{\theta(z)}{\phi(z)} = \sum_{j=-\infty}^{\infty} \psi_j z^j, \quad r^{-1} < |z| < r$$

para $r > 1$ (Expansión en series de Laurent). La región $r^{-1} < |z| < r$ es conocido como anillo.

Nota 3.20. Si los polinomios $\phi(z)$ y $\theta(z)$ no tienen ceros comunes entonces tenemos lo siguiente:

- Si las raíces de $\phi(z)$ están por fuera del círculo unitario, entonces la solución de la ecuación 3.1 es estacionaria y causal.
- Si algunas de las raíces de $\phi(z)$ están por fuera y otras por dentro del círculo unitario, entonces la solución de la ecuación 3.1 es estacionaria pero no es causal.
- Si las raíces de $\phi(z)$ están sobre del círculo unitario, entonces la solución de la ecuación 3.1 es no estacionaria.
- Qué sucede si los polinomios tiene ceros comunes? Ver Página 86 del libro [2].

Existe una definición análoga a la causalidad que es la invertibilidad. Veamos de que se trata.

Definición 3.21. Un proceso ARMA(p,q) definido por las ecuaciones $\phi(B)X_t = \theta(B)Z_t$ se llama invertible si existe una sucesión de constantes $\{\pi_j\}$ tal que $\sum_{j=0}^{\infty} |\pi_j| < \infty$ y

$$Z_t = \sum_{j=0}^{\infty} \pi_j X_{t-j},$$

para cada $t \in \mathbb{Z}$.

Teorema 3.22. *Sea $\{X_t\}$ un proceso ARMA(p,q) para el cual los polinomios $\phi(\cdot)$ y $\theta(\cdot)$ no tiene ceros comunes. Entonces $\{X_t\}$ es invertible si y sólo si $\theta(z) \neq 0$ para todo $z \in \mathbb{C}$ tal que $|z| \leq 1$, y los coeficientes son determinados por la relación*

$$\pi(z) = \sum_{j=0}^{\infty} \pi_j z^j = \frac{\phi(z)}{\theta(z)}$$

Ejercicio 3.23. *Cuál de los siguientes procesos son invertibles?*

- ◊ $X_t + 1.9X_{t-1} + 0.88X_{t-2} = Z_t + 0.2Z_{t-1} + 0.7Z_{t-2}$
- ◊ $X_t + 1.8X_{t-1} + 0.81X_{t-2} = Z_t + 0.8Z_{t-1}$

Nosotros fijaremos nuestra atención a procesos ARMA que son causables e invertibles, sin embargo los procesos ARMA no tiene por qué serlo. Para esto, entonces de resultados de la variable compleja sobre funciones analíticas racionales son necesarios para obtener el siguiente resultado.

Nota 3.24. *El proceso definido en 3.1 tiene media cero, sin embargo no siempre tiene que ser así. Un modelo ARMA con media diferente de cero se define de la siguiente manera:*

$$X_t = c + \phi_1 X_{t-1} + \phi_2 X_{t-2} + \cdots + \phi_p X_{t-p} + Z_t + \theta_1 Z_{t-1} + \cdots + \theta_q Z_{t-q},$$

donde $c \neq 0$, es decir, un modelo ARMA con intercepto. Entonces el valor esperado de éste proceso se puede verificar que es

$$E[X_t] = \frac{c}{1 - \phi_1 - \cdots - \phi_p}.$$

Sin embargo, ninguna de las propiedades anteriores se ven modificadas.

3.4 Forma de Computar la Función de Autocovarianza de un proceso ARMA

Dado que la F.A.C.V teórica de un proceso ARMA(p,q) causable, en forma general requiere del cálculo de una suma infinita, desde el punto de vista práctico resulta de forma ineficiente.

Métodos para el cómputo:

Método 1. La función de autocovarianza γ de un proceso ARMA(p,q) causal $\phi(B)X_t = \theta(B)Z_t$ entonces

$$\gamma(k) = \sigma^2 \sum_{j=0}^{\infty} \psi_j \psi_{j+|k|}$$

$$\text{con } \psi(z) = \sum \psi_j z^j = \frac{\theta(z)}{\phi(z)} \text{ para } |z| \leq 1.$$

$$\theta(z) = 1 + \theta_1 z + \cdots + \theta_q z^q$$

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

$$\gamma(z) = \frac{\theta(z)}{\phi(z)} \Rightarrow \psi(z)\phi(z) = \theta(z) \quad |z| \leq 1$$

como $\theta_0 = 1$ y $\theta_j = 0$ para $j > q$ y $\phi_j = 0$ para $j > p$ se obtiene:

$$(*) = \begin{cases} \psi_j - \sum_{0 < k \leq j} \phi_k \psi_{j-k} = \theta_j & \text{si } 0 \leq j < \max(p, q+1) \\ \psi_j - \sum_{0 < k \leq p} \phi_k \psi_{j-k} = 0 & \text{si } j \geq \max(p, q+1) \end{cases}$$

y los ψ 's se obtienen de solucionar esas ecuaciones. Se puede ver como una ecuación en diferencias en las ψ 's de forma sucesiva:

$$\psi_0 = \theta_0 = 1$$

$$\psi_1 = \theta_1 + \psi_0 \phi_1 = \theta_1 + \phi_1$$

$$\psi_2 = \theta_2 + \psi_0 \phi_2 + \phi_1 \theta_1 = \theta_2 + \phi_2 + \theta_1 \phi_1 + \phi_1^2$$

:

$$\psi_n = \sum_{i=1}^k \sum_{j=0}^{r_i-1} \alpha_{ij} n^j \xi_i^{-n}$$

donde $n \geq \max(p, q+1)$, ξ_i los ceros diferentes de $\phi(z)$ para $i = 1, 2, \dots, k$, r_i sus multiplicidades y los α_{ij} que se obtienen de las condiciones iniciales.

Ejercicio 3.25. Considera el proceso $(1 - B + \frac{1}{4}B^2)X_t = (1 + B)Z_t$,

entonces tenemos $\phi(z) = 1 - \underbrace{z}_{\phi_1} + \underbrace{\left(-\frac{1}{4}\right)}_{\phi_2}$ y $\theta(z) = 1 + \underbrace{z}_{\theta_1}$

$$\phi(z) = 0$$

$$1 - z + \frac{1}{4}z^2 = 0 \quad \phi_1 = 1$$

$$4 - 4z + z^2 = 0 \quad \phi_2 = -\frac{1}{4}$$

$$(z - 2)^2 = 0 \quad \theta_1 = 1$$

$z = 2$, es decir $\varepsilon_1 = 2$ de multiplicidad $r_1 = 1$.

Para $0 \leq j < \max(2, 2) = 2$, es decir, $0 \leq j < 2$ tenemos que $\psi_0 = \theta_0 = 1$

$$\psi_1 = \theta_1 + \psi_0\phi_1 = \theta_1 + \phi_1 = 2$$

para $j \geq 2$

$$\psi_2 - 1\psi_1 + \frac{1}{4}\psi_0 = 0 \text{ en general}$$

$$\underbrace{\psi_j - \psi_{j-1} + \frac{1}{4}\psi_{j-2}}_{\text{polinomio } \phi(z)} = 0$$

Así, la solución general tiene la forma:

$$\psi_n = \sum_{i=1}^k \sum_{j=0}^{r_i-1} \alpha_{ij} n^j \varepsilon_i^{-n}$$

con $k = 1$ y $r_1 = 2$.

$$\begin{aligned} \psi_n &= \sum_{i=1}^1 \sum_{j=0}^1 \alpha_{ij} n^j \varepsilon_i^{-n} = (\alpha_{10} n^0 2^{-n} + \alpha_{11} n^1 2^{-n}) \\ &= (\alpha_{10} + \alpha_{11} n) 2^{-n} \end{aligned}$$

donde α_{10} y α_{11} se obtienen de resolver:

para $n = 0$ tenemos que (3.7)

$$1 = \psi_0 = (\alpha_{10} + \alpha_{11} \cdot 0) 2^{-0} = \alpha_{10} \Rightarrow \alpha_{10} = 1$$

para $n = 1$ tenemos que (3.9)

$$2 = \psi_1 = (\alpha_{10} + \alpha_{11}) 2^{-1} \quad (3.10)$$

$$2 \cdot 2 = \alpha_{10} + \alpha_{11} \quad (3.11)$$

Solución

$$4 = \alpha_{10} + \alpha_{11} \Rightarrow \alpha_{11} = 3$$

Así

$$\psi_n = (1 + 3n) 2^{-n} \quad n = 0, 1, 2, \dots$$

Por lo tanto, la función de autocovarianza queda de la siguiente forma

3.4. FORMA DE COMPUTAR LA FUNCIÓN DE AUTOCOVARIANZA DE UN PROCESO

$$k \geq 0$$

$$\psi(k) = \sigma^2 \sum_{j=0}^{\infty} (1+3j) 2^{-j} (1+3(j+k)) 2^{-j-k} \quad (3.12)$$

$$= \sigma^2 \sum_{j=0}^{\infty} (1+3j)(1+3j+3k) 2^{-2j-k} \quad (3.13)$$

$$= \sigma^2 2^{-k} \sum_{j=0}^{\infty} (1+3j+3k+3j+9j^2+9jk) 2^{-2j} \quad (3.14)$$

$$= \sigma^2 2^{-k} \sum_{j=0}^{\infty} [(3k+1)4^{-j} + 3(2j+3k)j4^{-j} + 9j^24^{-j}] \quad (3.15)$$

$$= \sigma^2 2^{-k} \left[(3k+1) \frac{1}{1-\frac{1}{4}} + 3(3k+2) \frac{4}{9} + 9 \frac{20}{27} \right] \quad (3.16)$$

$$= \sigma^2 2^{-k} \left[\frac{32}{3} + 8k \right]$$

Método 2 . La idea de éste método consiste en tomar la ecuación en diferencias del proceso ARMA(p,q), multiplicar a ambos lados de la igualdad por X_{t-k} , y tomar valor esperado

$$\begin{aligned} E[X_t X_{t-k}] - \phi_1 E[X_{t-1} X_{t-k}] - \phi_2 E[X_{t-2} X_{t-k}] - \cdots - \phi_p E[X_{t-p} X_{t-k}] \\ = E[Z_t X_{t-k}] + \theta_1 E[Z_{t-1} X_{t-k}] + \cdots + \theta_q E[Z_{t-q} X_{t-k}]. \end{aligned}$$

Se puede verificar que las condiciones iniciales son:

$$\gamma(k) - \phi_1 \gamma(k-1) - \cdots - \phi_p \gamma(k-p) = \sigma^2 \sum_{k \leq j \leq q} \theta_j \psi_{j-k}, \quad 0 \leq k < \max(p, q+1)$$

y en forma general tenemos que

$$\gamma(k) - \phi_1 \gamma(k-1) - \cdots - \phi_p \gamma(k-p) = 0, \quad k > \max(p, q+1).$$

La última expresión se puede ver como una ecuación en diferencias homogénea no estocástica para $\gamma(k)$, cuya solución general es dada de manera análoga que para la ecuación en diferencias para ψ_n del método anterior:

$$\gamma(h) = \sum_{i=1}^k \sum_{j=0}^{r_i-1} \beta_{ij} h^j \varepsilon_i^{-h}, \quad h > \max(p, q+1) - p.$$

Note que las condiciones iniciales son necesarias para encontrar a los $\beta'_{ij}s$. Ver ejemplo 3.3.2 del libro [2].

Método 3. Este método no encuentra la expresión exacta de $\gamma(h)$, sino que lo hace de forma numérica usando la forma recurrente. La idea es usar primero

$$\gamma(k) - \phi_1\gamma(k-1) - \cdots - \phi_2\gamma(k-p) = \sigma^2 \sum_{k \leq j \leq q} \theta_j \psi_{j-k}, \quad 0 \leq k < \max(p, q+1)$$

y solucionar el sistema de ecuaciones para $k = 0, 1, \dots, p$. Las incógnitas por supuesto son $\gamma(0), \dots, \gamma(\max(p, q+1) - 1)$.

Después usamos

$$\gamma(k) - \phi_1\gamma(k-1) - \cdots - \phi_2\gamma(k-p) = 0, \quad k > \max(p, q+1),$$

para determinar $\gamma(\max(p, q+1)), \gamma(\max(p, q+1)+1), \dots$. Ver ejemplo 3.3.6 del libro [2].

Método 4. Se usa la función generadora de autocovarianza. La función generadora de autocovarianza para un proceso estocástico estacionario $\{X_t\}$ es definida como

$$G(z) = \sum_{k=-\infty}^{\infty} \gamma(k)z^k, \quad (3.17)$$

dado que la serie converge para todo $z \in \mathbb{C}$ en el anillo $r^{-1} < |z| < r$ con $r > 1$. La idea consiste en computar $G(z) = \sigma^2 \frac{\theta(z)\theta(z^{-1})}{\phi(z)\phi(z^{-1})}$ usando los polinomios del proceso ARMA(p,q), y usando la transformada z. Después se identifica cada coeficiente de z^k o z^{-k} $G(z)$ con $\gamma(k)$. Ver ejemplos sección 3.5 del libro [2].

3.5 Predicción de Procesos Estacionarios

El problema de obtener el pronóstico de los valores $\{X_t, t \geq n+1\}$ de un proceso estacionario en términos de $\{X_1, \dots, X_n\}$, consiste en predecir como es el comportamiento subsecuente de $\{X_t\}$ cuando se han tomado observaciones X_1, \dots, X_n . La solución en el espacio $L^2(\Omega, F, P) = L^2$ está basado en el teorema de la proyección ortogonal.

Planteemos el problema general de predicción en L^2 desde un enfoque matemático. El problema predecir a $X \in L^2$ con base en $Y_1, \dots, Y_k \in L^2$ consiste en encontrar una función medible Y con respecto de Y_1, \dots, Y_k , tal que la distancia de X a Y , es decir, $\|X - Y\|^2 = E[(X - Y)^2]$ (error cuadrático medio de predicción) sea mínima. Si definimos que $\mathcal{M}(Y_1, \dots, Y_k)$ es el subespacio

cerrado de L^2 que consiste de todas las variables aleatorias en L^2 de la forma $\phi(Y_1, \dots, Y_k)$ para alguna función de Borel(medible) $\phi : \mathbb{R}^k \rightarrow \mathbb{R}$. Entonces podemos usar el Teorema de Proyección en espacios de Hilbert.

Teorema 3.26. (Teorema de la Proyección) Si \mathcal{M} es un subespacio cerrado del espacio de Hilbert \mathcal{H} , y $x \in \mathcal{H}$, entonces

(i) existe un único elemento $\hat{x} \in \mathcal{M}$ tal que

$$\|x - \hat{x}\| = \inf_{y \in \mathcal{M}} \|x - y\|,$$

y

(ii) $\hat{x} \in \mathcal{M}$ y $\|x - \hat{x}\| = \inf_{y \in \mathcal{M}} \|x - y\|$ si y solo si $\hat{x} \in \mathcal{M}$ y $(x - \hat{x}) \in \mathcal{M}^\perp$.

Al elemento \hat{x} se le llama proyección ortogonal de x sobre \mathcal{M} .

Nota 3.27. A la función $P_{\mathcal{M}} : \mathcal{H} \rightarrow \mathcal{M}$ es llamado la función proyección de \mathcal{H} sobre \mathcal{M} . Denotaremos a la proyección $\hat{x} = P_{\mathcal{M}}x$, la cual se le conoce como la mejor predicción de x . Ver propiedades de la función proyección en libro[2] proposición 2.3.2. Retornando al problema de predicción en L^2 planteado, entonces usando el teorema de la proyección tenemos que la mejor función medible Y de Y_1, \dots, Y_k es $P_{\mathcal{M}(Y_1, \dots, Y_k)}X$, la cual es evidente que es $E[X|Y_1, \dots, Y_k]$ directamente de la definición de esperanza condicional. Directamente del teorema de la proyección se tiene que el elemento de \mathcal{M} mas cercano a x es el único elemento $\hat{x} \in \mathcal{M}$ tal que

$$\langle x - \hat{x}, y \rangle = 0, \quad \text{para todo } y \in \mathcal{M}. \quad (3.18)$$

A las ecuaciones 3.18 se les llama **ecuaciones de predicción**, y son las que permiten encontrar la proyección ortogonal \hat{x} al resolver el sistema.

Nota 3.28. Sólo en algunos casos especiales, el mejor predictor $P_{\mathcal{M}(Y_1, \dots, Y_k)}X$ puede ser encontrado en forma analítica ya que en general es muy difícil. Por lo tanto, es muy frecuente que se restrinja la atención al predictor sobre el subespacio, llamado subespacio generado $\overline{SP}\{Y_1, \dots, Y_k\}$, que consiste sobre el conjunto de todas las combinaciones lineales de la forma $\alpha_1Y_1 + \dots + \alpha_kY_k$, donde los $\alpha_1, \dots, \alpha_k \in \mathbb{R}$. Así, $P_{\overline{SP}\{1, Y_1, \dots, Y_n\}}X = \alpha_1Y_1 + \dots + \alpha_kY_k$, y las constantes $\alpha_1Y_1 + \dots + \alpha_kY_k$ deben ser encontradas resolviendo las ecuaciones de predicción 3.18., lo cual nos lleva a resolver el sistema de ecuaciones lineales

$$\langle X - P_{\overline{SP}\{1, Y_1, \dots, Y_n\}}X, Y \rangle = 0, \quad \text{para todo } Y \in \overline{SP}\{1, Y_1, \dots, Y_n\}$$

lo cual es equivalente

$$\langle P_{\overline{SP}\{1, Y_1, \dots, Y_n\}} X, Y_j \rangle = \langle X, Y_j \rangle, \quad j = 1, \dots, k$$

y así usando la forma lineal tenemos

$$\sum_{i=1}^k \alpha_i \langle Y_i, Y_j \rangle = \langle X, Y_j \rangle, \quad j = 1, \dots, k.$$

Finalmente tenemos dos tipos de predictores que nos interesan para la predicción en series de tiempo con $Y = X_{n+h}$ $h = 1, 2, \dots$:

$$P_{\mathcal{M}\{1, X_1, \dots, X_n\}} Y \quad y \quad P_{\overline{SP}\{1, X_1, \dots, X_n\}} Y,$$

se conocen como el mejor predictor y el mejor predictor lineal de Y .

Vamos a enunciar algunas de las propiedades del predictor de una forma que sean aplicables mas directamente.

Proposición 3.29. Propiedades del operador $P(\cdot | \tilde{\mathbf{W}})$

Supongamos que las variables Y y W_n, \dots, W_1 son tales que tienen segundos momentos finitos y $\mu_Y = E[Y]$, $\mu_i = E[W_i]$ y covarianzas $Cov(Y, Y)$, $Cov(Y, W_i)$ y $Cov(W_i, W_j)$ conocidas. Sea $\tilde{\mathbf{W}} = (W_n, \dots, W_1)$ y $\mu_{\tilde{\mathbf{W}}} = (\mu_n, \dots, \mu_1)$, $\gamma = Cov(Y, \tilde{\mathbf{W}})$, $\Gamma = Cov(\tilde{\mathbf{W}}, \tilde{\mathbf{W}})$, entonces tenemos que el mejor predictor lineal de Y en términos de $\{1, W_n, \dots, W_1\}$ es

$$P(Y | \tilde{\mathbf{W}}) = \mu_Y + \tilde{\mathbf{a}}' (\tilde{\mathbf{W}} - \mu_{\tilde{\mathbf{W}}}) \tilde{\mathbf{a}}$$

con $\tilde{\mathbf{a}}$ una solución de

$$\tilde{\Gamma} \tilde{\mathbf{a}} = \gamma,$$

y el error cuadrático medio del predictor es

$$E[\{Y - P(Y | \tilde{\mathbf{W}})\}^2] = Var(Y) - \tilde{\mathbf{a}}' \gamma.$$

Adicionalmente tenemos las siguientes propiedades:

Supongamos adicionalmente que $E[U^2] < \infty$ y $E[V^2] < \infty$, Γ como antes y $\beta, \alpha_1, \dots, \alpha_n$ constantes, entonces

1. $P(U | \tilde{\mathbf{W}}) = E[U] + \tilde{\mathbf{a}}' (\tilde{\mathbf{W}} - E[\tilde{\mathbf{W}}])$, donde $\tilde{\Gamma} \tilde{\mathbf{a}}' = Cov(U, \tilde{\mathbf{W}})$.
2. $E[(U - P(U | \tilde{\mathbf{W}})) \tilde{\mathbf{W}}] = \tilde{\mathbf{0}}$ y $E[U - P(U | \tilde{\mathbf{W}})]$.
3. $E[(U - P(U | \tilde{\mathbf{W}}))^2] = Var(U) - \tilde{\mathbf{a}}' Cov(U, \tilde{\mathbf{W}})$.

4. $P(\alpha_1 U + \alpha_2 V + \beta | \tilde{\mathbf{W}}) = \alpha_1 P(U | \tilde{\mathbf{W}}) + \alpha_2 P(V | \tilde{\mathbf{W}}) + \beta.$
5. $P(\sum_{i=1}^n \alpha_i W_i + \beta | \tilde{\mathbf{W}}) = \sum_{i=1}^n \alpha_i W_i + \beta.$
6. $P(U | \tilde{\mathbf{W}}) = E[U] \text{ si } Cov(U, \tilde{\mathbf{W}}) = 0.$
7. $P(U | \tilde{\mathbf{W}}) = P\left(P(U | \tilde{\mathbf{W}}, \tilde{\mathbf{V}}) | \tilde{\mathbf{W}}\right).$

Ejemplo 3.30. Considere que el proceso $\{X_t\}$ es un AR(1) definido como

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

con $|\phi| < 1$ y $\{Z_t\} \sim RB(0, \sigma^2)$.

Además, dado que $|\phi| < 1$ entonces tenemos que $X_t = \sum_{j=0}^{\infty} \phi^j Z_{t-j}$, entonces Z_{t+s} es no correlacionado con X_t , es decir, es ortogonal a X_t para todo $s > 0$.

Supongamos que deseamos predecir X_{n+1} en términos de $\{X_n, \dots, X_1\}$ usando el predictor lineal, es decir

$$\hat{X}_{n+1|n} = P_{\overline{SP}\{1, X_1, \dots, X_n\}} X_{n+1}.$$

Usando la ecuación del proceso AR(1), tenemos que

$$\hat{X}_{n+1|n} = P_{\overline{SP}\{1, X_1, \dots, X_n\}} X_{n+1} = P_{\overline{SP}\{1, X_1, \dots, X_n\}} (\phi X_n + Z_{n+1}),$$

usando la propiedad de linealidad de la proyección ortogonal, tenemos

$$P_{\overline{SP}\{1, X_1, \dots, X_n\}} (\phi X_n) + P_{\overline{SP}\{1, X_1, \dots, X_n\}} (Z_{n+1}).$$

Ahora, usando la propiedad de ortogonalidad o la no autocorrelación del proceso $\{Z_t\}$, con respecto al proceso $\{X_t\}$ cuando el proceso de ruido está en el futuro, tenemos que $P_{\overline{SP}\{1, X_1, \dots, X_n\}} (Z_{n+1}) = E[Z_{n+1}] = 0$. Así

$$\hat{X}_{n+1|n} = P_{\overline{SP}\{1, X_1, \dots, X_n\}} (\phi X_n),$$

con lo cual usando la propiedad (i)(linealidad del operador proyección) y (v)($x \in \mathcal{M}$ si y sólo si $P_{\mathcal{M}}x = x$) de la proposición(2.3.2) en el libro [2], tenemos que $P_{\overline{SP}\{1, X_1, \dots, X_n\}} (\phi X_n) = X_n$, con lo cual,

$$\hat{X}_{n+1|n} = \phi X_n.$$

Por otro lado, no es difícil verificar que

$$ECM(X_{n+1}) = E[(X_{n+1} - \hat{X}_{n+1|n})^1] = \sigma_z^2.$$

Verifique también que :

$$E[X_{n+1} - \hat{X}_{n+1|n}] = 0,$$

es decir, el predictor es insesgado.

Cómo se obtiene analíticamente el predictor lineal dos pasos adelante? cómo se computa su incertidumbre?

3.5.1 Las ecuaciones de predicción en el dominio del tiempo

Sea $\{X_t\}$ un proceso estocástico estacionario de media μ y FACV $\gamma(\cdot)$. Entonces, el proceso $\{Y_t\} = \{X_t - \mu\}$ es un proceso estacionario de media cero con FACV $\gamma(\cdot)$ y se puede verificar [Tarea](#)

$$P_{\overline{SP}\{1, X_1, \dots, X_n\}} X_{n+h} = \mu + P_{\overline{SP}\{Y_1, \dots, Y_n\}} Y_{n+h}.$$

Vamos asumir que $\mu = 0$ y así

$$P_{\overline{SP}\{1, X_1, \dots, X_n\}} X_{n+h} = P_{\overline{SP}\{Y_1, \dots, Y_n\}} Y_{n+h}.$$

Las ecuaciones de predicción un paso adelante

Sea $H_n = \overline{SP}(1, X_1, \dots, X_n)$ el subespacio lineal cerrado, $n \geq 1$ y sea \hat{X}_{n+1} los predictores un paso adelante definidos como

$$\hat{X}_{n+1} = \begin{cases} 0 & \text{si } n = 0 \\ P_{H_n} X_{n+1} & n \geq 1 \end{cases}$$

Es decir, para $n \geq 1$

$$\hat{X}_2 = P_{\overline{SP}\{X_1\}} X_2 \tag{3.19}$$

$$\hat{X}_3 = P_{\overline{SP}\{X_1, X_2\}} X_3 \tag{3.20}$$

$$\hat{X}_4 = P_{\overline{SP}\{X_1, X_2, X_3\}} X_4 \tag{3.21}$$

Puesto que $\hat{X}_{n+1} \in H_n$, $n \geq 1$, entonces se puede escribir $\hat{X}_{n+1} = \phi_{n1}X_n + \phi_{n2}X_{n-1} + \dots + \phi_{nn}X_1$ $n \geq 1$ tal que $\phi_{n1}, \dots, \phi_{nn}$ satisfacen las ecuaciones de predicción

$$\left\langle \sum_{i=1}^n \phi_{ni} X_{n+1-i}, X_{n+1-j} \right\rangle = \langle X_{n+1}, X_{n+1-j} \rangle \quad j = 1, 2, \dots, n$$

Como $\langle X, Y \rangle = E[XY]$ y por linealidad del producto interno, tenemos que

$$\sum_{i=1}^n \phi_{ni} \gamma(i-j) = \gamma(j) \quad j = 1, 2, \dots, n$$

lo cual es equivalente a

$$\Gamma_n \phi_n = \gamma_n (*) \quad \text{Ecuación de predicción}$$

y así,

$$\phi_n = \Gamma_n^+ \gamma_n,$$

con

$$\Gamma_n = [\gamma(i-j)]_{i,j=1,2,\dots,n} \quad \phi_n = (\phi_{n1}, \dots, \phi_{nn})' \quad \gamma_n = (\rho_{(1)}, \dots, \rho_{(n)})'$$

El teorema de la predicción garantiza que (*) tiene al menos una solución que define unicidad con unicidad a \hat{X}_{n+1} . Veamos el siguiente resultado el cual son las condiciones que garantizan la no singularidad de Γ_n y por lo tanto su unicidad.

Proposición 3.31. *Si $\gamma(0) > 0$ y $\gamma(h) \rightarrow 0$ cuando $h \rightarrow \infty$ entonces la matriz de covarianzas $\Gamma_n = [\gamma(i-j)]_{i,j=1,2,\dots,n}$ de $(X_1, \dots, X_n)'$ es no singular para cada h .*

Lema 3.32. *Bajo las condiciones del resultado anterior, el mejor predictor lineal \hat{X}_{n+1} de X_n en términos de X_1, \dots, X_n es:*

$$\hat{X}_{n+1} = \sum_{i=1}^n \phi_{ni} X_{n+1-i}, \quad n = 1, 2, \dots$$

donde $\phi_n = (\phi_{n1}, \dots, \phi_{nn})' = \Gamma_n^{-1} \gamma_n$, $\gamma_n = (\gamma(1), \dots, \gamma(n))'$ y $\Gamma_n = [\gamma(i-j)]$ $i, j = 1, \dots, n$.

El error cuadrático medio de predicción es $v_n = \gamma(0) - \gamma_n \Gamma_n \gamma_n$

$$v_n = E[(X_{n+1} - \hat{X}_{n+1})^2]$$

$$\begin{aligned}\hat{X}_{n+1} &= \phi'_n X_n & X_n &= (X_1, \dots, X_n)' \\ &= \Gamma_n^{-1} \gamma_n\end{aligned}\tag{3.22}$$

$$v_n = E[(X_{n+1} - \phi'_n X_n)^2] \tag{3.23}$$

$$= E[(X_{n+1} - (\Gamma_n^{-1} \gamma_n)' X_n)^2] \tag{3.24}$$

$$= E[X_{n+1}^2] - 2E[X_{n+1}(\Gamma_n^{-1} \gamma_n) X_n] + E[(\Gamma_n^{-1} \gamma_n)' X_n]^2 \tag{3.25}$$

$$= \gamma(0) - 2\gamma'_n \Gamma_n^{-1} \gamma_n + \gamma'_n \Gamma_n^{-1} \gamma_n \tag{3.26}$$

$$= \gamma(0) - \gamma'_n \Gamma_n^{-1} \gamma_n.$$

Ecuaciones para los predictores h-pasos adelante

El mejor predictor de X_{n+h} en términos de X_1, \dots, X_n para cualquier $h \geq 1$. Puede ser encontrado de la misma manera como \hat{X}_{n+1} . Así que

$$P_{H_n} X_{n+h} = \phi_{n1}^{(h)} X_n + \dots + \phi_{nn}^{(h)} X_1$$

donde $\phi_n^{(h)} = (\phi_{n1}^{(h)}, \dots, \phi_{nn}^{(h)})'$ es cualquier solución

$$\Gamma_n \phi_n^{(h)} = \gamma_n^{(h)}$$

$$\text{con } \gamma_n^{(h)} = (\gamma(h), \gamma(h+1), \dots, \gamma(n+h+1))'$$

Ejemplo 3.33. Considere que el proceso $\{X_t\}$ es un AR(1) definido como

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

con $|\phi| < 1$ y $\{Z_t\} \sim RB(0, \sigma^2)$.

Supongamos que deseamos predecir X_{n+1} en términos de $\{X_n, \dots, X_1\}$ usando el predictor lineal. Entonces el predictor puede escribirse como

$$P_{\overline{SP}\{X_n, \dots, X_1\}} X_{n+1} = P_n X_{n+1} = \underbrace{a'_n}_{} \underbrace{X_n}_{} \tag{*}$$

donde $\underbrace{X_n}_{} = (X_n, \dots, X_1)'$ y $\underbrace{a'_n}_{} \satisfies$

$$\underbrace{\Gamma_n a'_n}_{} = \gamma_n(1),$$

donde

$$\Gamma_n = [\gamma(i-j)]_{i,j=1}^n$$

y $\gamma_n(1) = (\gamma(1), \gamma(2), \dots, \gamma(n))'$. Para el caso AR(1) omitiendo el denominador ya que es común a todas las covarianzas, tenemos entonces

$$\Gamma_n = \begin{bmatrix} 1 & \phi & \phi^2 & \cdots & \phi^{n-1} \\ \phi & 1 & \phi & \cdots & \phi^{n-2} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ \phi^{n-1} & \phi^{n-2} & \phi^{n-3} & \cdots & 1 \end{bmatrix},$$

y $\gamma_n(1) = (\phi, \phi^2, \dots, \phi^n)'$, donde una solución del sistema de ecuaciones es $\underset{\sim}{a}'_n = (\phi, 0, 0, \dots, 0)'$ ya que Γ_n es una matriz de Toeplitz. Así el mejor predictor de X_{n+1} en términos de $\{X_n, \dots, X_1\}$ es

$$\underset{\sim}{P}_n X_{n+1} = \underset{\sim}{a}'_n \underset{\sim}{X}_n = \phi X_n,$$

y el error cuadrático medio del predictor es

$$E[(X_{n+1} - \underset{\sim}{P}_n X_{n+1})^2] = \sigma^2.$$

Tarea: Cómo queda $P_n X_{n+2}$ y su error de cuadrático medio.? Cómo queda en general $P_n X_{n+h}$? Cómo queda la predicción un paso adelante para un modelo MA(1) para los casos particulares de $n=2$ y 3 ?

.

Nota 3.34. Note que al aumentar el horizonte de pronóstico, la incertidumbre aumenta.

Nota 3.35. En la práctica existen varios algoritmos que permiten obtener los pronósticos en forma iterativa: Durbin-Levinson, Innovaciones, Usando el Filtro de Kalman para modelos de Espacio-Estado.

Por ejemplo, uno ver que usando el algoritmo de innovaciones, se puede obtener una forma recursiva de como son las predicciones un paso adelante usando el algoritmo de innovaciones como sigue:

Vamos a considerar un proceso ARMA(p,q) causable

$$\phi(B)X_t = \theta(B)Z_t \quad \{Z_t\} \sim RB(0, \sigma^2)$$

Establezcamos el proceso $\{W_t\}$ de la siguiente manera

$$W_t = \sigma^{-1}X_t \quad t = 1, 2, \dots, m \quad (3.27)$$

$$W_t = \sigma^{-1}\phi(B)X_t \quad t > m \quad (3.28)$$

Donde $m = \max(p, q)$

Definamos $\theta_0 = 1$ y supongamos que $p, q \geq 1$.

$$H_n = \overline{SP}\{X_1, \dots, X_n\} = \overline{SP}\{W_1, \dots, W_n\}$$

Para $n > 1$ \hat{X}_{n+1} y \hat{W}_{n+1} son $P_{H_n}X_{n+1}$ y $P_{H_n}W_{n+1}$ respectivamente.

Definiendo $\hat{X}_1 = 0$. La función de autocovarianza $\gamma_X(\cdot)$ se puede computar asumiendo cualquier método antes mencionado. Se puede comprobar que $k(i, j) = E[W_i W_j]$ queda de la siguiente forma

$$k(i, j) = \begin{cases} \sigma^{-2} \gamma_X(i - j) & 1 \leq i, j \leq m \\ \sigma^{-2} [\rho(i - j) - \sum_{r=1}^p \phi_r \gamma_X(r - |i - j|)] & \min(i, j) \leq m \leq \max(i, j) \leq 2m \\ \sum_{r=0}^q \theta_r \theta_{r+|i-j|} & \min(i, j) > m \\ 0 & \text{en otro caso} \end{cases}$$

estableciendo que $\theta_j = 0$ para $j > q$

Aplicando el algoritmo de innovación al proceso $\{W_t\}$ tenemos que

$$\begin{aligned} \hat{W}_{n+1} &= \sum_{j=1}^n \theta_{nj} (W_{n+1-j} - \hat{W}_{n+q-j}) & 1 \leq n < m \\ \hat{W}_{n+1} &= \sum_{j=1}^q \theta_{nj} (W_{n+1-j} - \hat{W}_{n+1-j}) & n \leq m. \end{aligned} \quad (3.29)$$

Para encontrar \hat{X}_n de \hat{W}_n , podemos aplicar a ambos lados proyección ortogonal sobre H_{t-1}

$$\begin{cases} \hat{W}_t = \sigma^{-1} \hat{X}_t & t = 1, 2, \dots, m \\ \hat{W}_t = \sigma^{-1} [\hat{X}_t - \phi_1 X_{t-1} - \dots - \phi_p X_{t-p}] & t > m, \end{cases}$$

se puede verificar así que

$$X_t - \hat{X}_t = \sigma(W_t - \hat{W}_t) \quad t \geq 1$$

Reemplazando $W_j - \hat{W}_j$ por $\sigma^{-1}(X_j - \hat{X}_j)$ en 3.29 anteriormente tenemos que

$$\begin{cases} \hat{X}_{n+1} = \sum_{j=1}^n \theta_{nj} (X_{n+1-j} - \hat{X}_{n+1-j}) & 1 \leq n \leq m \\ \hat{X}_{n+1} = \phi_1 X_n + \dots + \phi_p X_{n+1} + \sum_{j=1}^p \theta_{nj} (X_{n+1-j} - \hat{X}_{n+1-j}) & n \geq m \end{cases}$$

y $E[(X_{n+1} - \hat{X}_{n+1})^2] = \sigma^2 r_n$. Hay una expresión similar para los predictores h-pasos adelante, ver [2] Página 179.

Predicción de un proceso Gaussiano (Bandas de Predicción)

Sea $\{X_t\}$ un proceso Gaussiano estacionario de media cero y FACV $\gamma(\cdot)$ tal que $\gamma(0) > 0$ y $\gamma(0) \rightarrow \infty$ $h \rightarrow +\infty$.

Se sabe que

$$P_n X_{n+h} = \Gamma_n^{-1} [\gamma_{n+h-1}, \gamma_{n+h-2}, \dots, \gamma_n] X_n \quad h \geq 1.$$

Bajo Gaussianidad se puede verificar que

$$P_{\overline{SP}, \{1, X_1, \dots, X_n\}} X = E_{\mu(X_1, \dots, X_n)} X,$$

con lo cual

$$P_n X_{n+h} = E_{\mu(Z_1, \dots, Z_n)} X_{n+h} = E[X_{n+h} | X_1, \dots, X_n].$$

Por supuesto

$$\begin{aligned} \Delta_n(h) &= X_{n+h} - P_n X_{n+h} \sim N(0, \sigma_n^2(n)) \\ \sigma_n^2(h) &= E(\Delta_n^2(h)). \end{aligned} \tag{3.30}$$

Con lo cual un intervalo de predicción del $100(1 - \alpha)\%$ para X_{n+h} , es

$$P_n X_{n+h} \pm z_{1-\alpha/2} \sigma_n(h)$$

Que sucede si $\{X_t\}$ NO es Gaussiano?

Intervalos de Predicción

Esta sección la pueden estudiar en el libro [29] capítulo 11, también puede leer secciones 5.5 y 5.9 del libro [19].

Distribución Predictiva- Pronósticos Probabilísticos

Hoy en día, el pronóstico puntual, inclusive su respectivo intervalo de predicción resulta insuficiente, es por esto que se requiere conocer y estudiar la distribución de pronóstico o distribución predictiva, es decir, conocer

$$p(x_{T+h} | x_T, x_{T-1}, \dots, x_1),$$

es decir la distribución condicional de $X_{T+h} | X_T, X_{T-1}, \dots, X_1$. Note que la esperanza condicional, que es la mejor predicción, puede obtenerse de la distribución predictiva.

3.6 Función de Autocorrelación Parcial

Al igual que la FACP la FACV nos da información importante acerca de la estructura de autocorrelación de un proceso específico. La autocorrelación $\alpha(k)$ en el rezago k es la correlación entre X_1 y X_{k+1} ajustado por las intervenciones de las observaciones X_2, \dots, X_k .

Definición 3.36. *La función de autocorrelación parcial FACP $\alpha(\cdot)$ de un proceso estocástico estacionario*

$$\alpha(1) = \text{Corr}(X_2, X_1) = \rho(1)$$

y

$$\alpha(k) = \text{corr}(X_{k+1} - P_{\overline{SP}\{\{1, X_2, \dots, X_k\}} X_{k+1}}, X_1 - P_{\overline{SP}\{\{1, X_2, \dots, X_k\}} X_1})$$

donde la proyección $P_{\overline{SP}\{\{1, X_2, \dots, X_k\}} X_{k+1}}$ y $P_{\overline{SP}\{\{1, X_2, \dots, X_k\}} X_1}$ se pueden encontrar usando las ecuaciones de predicción. $\alpha(k)$ entonces es la autocorrelación parcial de rezago k .

Nota 3.37. *La autocorrelación parcial $\alpha(k)$ es la correlación de los dos residuales después de regresar X_{k+1} y X_1 sobre las observaciones intermedias X_2, \dots, X_k .*

Ejercicio 3.38. *Sea X_t un proceso AR(1) definido como:*

$$X_t = 0.9X_{t-1} + Z_t$$

$$\alpha(1) = \text{corr}(X_2, X_1) \tag{3.31}$$

$$= \text{Corr}(0.9X_1 + Z_2, X_1) \tag{3.32}$$

$$= 0.9 \quad \text{ya que } \text{corr}(Z_2, X_1) = 0$$

se puede ver que $P_{SP(1, X_2, \dots, X_k)} X_{k+1} = 0.9X_k$ usando el problema 2.1.2 y $P_{SP(X_2, \dots, X_k)} X_1 = 0.9X_2$ ya que (X_1, X_2, \dots, X_k) tiene la misma matriz de covarianza que $(X_{k+1}, X_k, \dots, X_2)$ tiene la misma solución. Para los α'_i s. En general, para cualquier $k \geq 2$.

$$\alpha(k) = \text{Corr}(X_{k+1} - 0.9X_k, X_1 - 0.9X_2) \tag{3.33}$$

$$= \text{Corr}(Z_{k+1}, X_1 - 0.9X_2) \tag{3.34}$$

$$= 0$$

Ejercicio 3.39. El proceso $MA(1)$

$$X_t + Z_t + \theta Z_{t-1}$$

$$\text{Para un } MA(1) \text{ se tiene: } Cov(X_{t+h}, X_t) = \begin{cases} \sigma^2 \sum_{j=0}^{1-|h|} \theta_j \theta_{j+|h|} & \text{si } |h| \leq 1 \\ 0 & \text{si } |h| > 1 \end{cases}$$

$$\text{Si } h = 1 \Rightarrow Cov(X_{t+1}, X_t) = \sigma^2 \theta_0 \theta = \sigma^2 \theta_1$$

$$\text{Si } h = 0 \Rightarrow Cov(X_t, X_t) = (1 + \theta_1^2)$$

Por lo tanto:

$$\alpha(1) = \rho(1) = \frac{\theta}{1+\theta^2}$$

$$\alpha(2) = Corr(X_3 - P_{SP(X_2)}X_3, X_1 - P_{SP(X_2)}X_1) P_{SP(X_2)}X_3 = \alpha X_2$$

$$< X_3 - \alpha X_2, X_2 > = 0$$

$$< X_3, X_2 > = < \alpha X_2, X_2 >$$

$$\sigma^2 \theta_1 = \alpha \sigma^2 (1 + \theta_1^2)$$

$$\Rightarrow \alpha = \frac{\theta}{1+\theta_1^2} \quad P_{SP(X_2)}X_3 = \frac{\theta}{1+\theta_1^2} X_2 \quad P_{SP(X_2)}X_1 = \beta X_2$$

$$< X_1 - \beta X_2, X_2 > = 0$$

$$< X_1, X_2 > = \beta < X_2, X_2 >$$

$$\beta = \frac{\theta_0}{1+\theta_1^2} \Rightarrow P_{SP(X_2)}X_1 = \frac{\theta_1}{1+\theta_1^2}$$

Con lo cual

$$\alpha(2) = \text{corr} \left(X_3 - \frac{\theta}{1+\theta^2} X_2, X_1 - \frac{\theta}{1+\theta^2} X_2 \right) \quad (3.35)$$

$$= \text{corr} \left(\frac{(1+\theta^2)Z_3 + \theta^3 Z_2 - \theta^2 Z_1}{1+\theta^2}, \frac{Z_1 + \theta(1+\theta^2)Z_0 - \theta Z_2}{1+\theta^2} \right) \quad (3.36)$$

$$= \frac{\frac{-\theta^4 \sigma^2 - \theta^2 \sigma^2}{(1+\theta^2)^2}}{\sqrt{\frac{\sigma^2}{(1+\theta^2)^2}} ((1+\theta^2)^2 + \theta^6 + \theta^4) \frac{\sigma^2}{(1+\theta^2)^2} (1+\theta^2)(1+\theta^2)^2 + \theta^2)} \quad (3.37)$$

$$= \frac{\frac{-\theta^2(1+\theta^2)}{(1+\theta^2)^2}}{\sqrt{(1+2\theta^2+\theta^4+\theta^6)(1+\theta^2+2\theta^4+\theta^6+2\theta^2)}} \quad (3.38)$$

$$= \frac{-\theta^2(1+\theta^2)}{\sqrt{(1+2\theta^2+2\theta^4+\theta^6)(1+2\theta^2+2\theta^4+\theta^6)}} \quad (3.39)$$

$$= \frac{-\theta^2(1+\theta^2)}{1+2\theta^2+2\theta^4+\theta^6} \quad (3.40)$$

$$= \frac{-\theta^2(1+\theta^2)}{(1+\theta^2)(1+\theta^2+\theta^4)} \quad (3.41)$$

Solución

$$= \frac{-\theta^2}{1+\theta^2+\theta^4}$$

En general se puede ver que (Usando un resultado que veremos , adelante)

$$\alpha(k) = \frac{(-\theta)^k(1-\theta^2)}{1-\theta^{2(k+1)}}$$

Ejemplo 3.40. Sea X_t un proceso AR(p) causable

$$X_t - \phi_1 X_{t-1} + \cdots + \phi_p X_{t-p} = Z_t \quad Z_t \sim RB(0, \sigma^2)$$

para $k > p$

$P_{\overline{SP}(X_2, \dots, X_k)} X_{k+1} = \sum_{j=1}^p \phi_j X_{k+1-j}$, puesto que si $Y \in \overline{SP}(X_2, \dots, X_K)$ entonces por causalidad $Y \in \overline{SP}(Z_j, j \geq k)$ y así,

$$< X_{k+1} - \sum_{j=1}^p \phi_j X_{k+1-j}, Y > = < Z_{k+1}, Y > = 0$$

para $k > p$ se concluye que :

$$\alpha(k) = \text{corr}(X_{k+1} - \sum_{j=1}^p \phi_j X_{k+1-j}, X_1 - P_{\overline{SP}(X_2, \dots, X_k)}) \quad (3.42)$$

$$= \text{corr}(Z_{k+1}, X_1 - P_{X_2, \dots, X_k} X_1) = 0 \quad (3.43)$$

Para $k \leq p$ los valores de $\alpha(k)$ se puede computar de forma fácil de la forma definición equivalente.

Definición 3.41. La autocorrelación parcial $\alpha(k)$ de $\{X_t\}$, un proceso estocástico estacionario de media cero y función de autocovarianza $\gamma(\cdot)$ tal que $\gamma(h) \rightarrow 0$ cuando $h \rightarrow \infty$ se define como

$$\alpha(k) = \phi_{kk} \quad k \geq 1$$

donde ϕ_{kk} son los coeficientes de la expresión

$$P_{\overline{SP}(X_1, \dots, X_k)} X_{k+1} = \sum_{j=1}^k \phi_{kj} X_{k+1-j}$$

los cuales son obtenidas resolviendo

$$\begin{bmatrix} \rho(0) & \rho(1) & \cdots & \rho(k-1) \\ \rho(1) & \rho(0) & \cdots & \rho(k-2) \\ \vdots & \vdots & \ddots & \vdots \\ \rho(k-1) & \rho(k-2) & \cdots & \rho(0) \end{bmatrix} \begin{bmatrix} \phi_{k1} \\ \phi_{k2} \\ \vdots \\ \phi_{kk} \end{bmatrix} = \begin{bmatrix} \rho(1) \\ \rho(2) \\ \vdots \\ \rho(k) \end{bmatrix}$$

Definición 3.42. La función de autocorrelación parcial muestral $\hat{\alpha}(k)$ en el rezago k de X_1, \dots, X_k se define, dado que $x_i \neq x_j$, para algún i, j por

$$\hat{\alpha}(k) = \hat{\phi}_{kk} \quad 0 \leq k < n.$$

4

Estimación de la media y funciones de autocovarianza autocorrelación parcial de un Proceso Estocástico Estacionario

En éste capítulo veremos como proceder a estimar y posteriormente a hacer inferencia para algunas cantidades de interés de un proceso estocástico estacionario, como lo son: la media y su función de autocorrelación simple. Adicionalmente, indicaremos como hacer inferencia acerca de la función de autocorrelación parcial. Note que sólo tenemos disponible una muestra X_1, \dots, X_T , es por esto que debemos suponer que el proceso es estacionario.

4.1 Estimación de la Media

Veamos como es la estimación de la media de un proceso estocástico estacionario. Un estimador insesgado natural de la media μ es

$$\bar{X}_n = \frac{1}{n}(X_1 + X_2 + \dots + X_n).$$

Veamos el comportamiento del error cuadrático medio, es decir, $E[(\bar{X}_n - \mu)^2]$ cuando $n \rightarrow \infty$.

Teorema 4.1. *Si $\{X_t\}$ es estacionario de media μ y FACV $\gamma(\cdot)$, entonces cuando $n \rightarrow \infty$*

$$Var(\bar{X}_n) = E[(\bar{X}_n - \mu)^2] \rightarrow 0 \quad \text{si} \quad \gamma(n) \rightarrow 0$$

y

$$nE[(\bar{X}_n - \mu)^2] \rightarrow \sum_{h=-\infty}^{\infty} \gamma(h) \quad \text{si} \quad \sum_{h=-\infty}^{\infty} |\gamma(h)| < \infty.$$

Demostración.

$$\begin{aligned} nVar(\bar{X}_n) &= \frac{1}{n} Var(X_1 + \cdots + X_n) \\ &= \frac{1}{n} Cov(X_1 + \cdots + X_n, X_1 + \cdots + X_n) \\ &= \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^n Cov(X_i, X_j) \end{aligned}$$

si $n = 2$

$$\begin{aligned} \frac{1}{2} \sum_{i,j=1}^2 Cov(X_i, X_j) &= \frac{1}{2} [Cov(X_1, X_1) + Cov(X_1, X_2) + Cov(X_2, X_1) + Cov(X_2, X_2)] \\ &= \frac{1}{2} [2\gamma(0) + \gamma(-1) + \gamma(1)] \end{aligned}$$

$n = 3$

$$\begin{aligned} \frac{1}{3} \sum_{i,j=1}^3 Cov(X_i, X_j) &= \frac{1}{3} [Cov(X_1, X_1) + Cov(X_1, X_2) + Cov(X_1, X_3) + Cov(X_2, X_1) + \\ &\quad Cov(X_2, X_2) + Cov(X_2, X_3) + Cov(X_3, X_1) + Cov(X_3, X_2) + Cov(X_3, X_3)] \\ &= \frac{1}{3} [\gamma(0) + \gamma(-1) + \gamma(-2) + \gamma(1) + \gamma(0) + \gamma(-1) + \gamma(2) + \gamma(1) + \gamma(0)] \\ &= \frac{1}{3} \sum_{h=-2}^2 (3 - |h|)\gamma(h) \end{aligned}$$

En general,

$$nVar(\bar{X}_n) = \frac{1}{n} \sum_{h=-(n-1)}^{n-1} (n - |h|)\gamma(h) = \sum_{|h|< n} \left(1 - \frac{|h|}{n}\right) \gamma(h).$$

Ahora, expandiendo la segunda suma tenemos

$$\begin{aligned}
 &= \gamma(0) + \underbrace{\left(1 - \frac{1}{n}\right)}_{<1} \gamma(1) + \underbrace{\left(1 - \frac{1}{n}\right)}_{<1} \gamma(-1) + \underbrace{\left(1 - \frac{2}{n}\right)}_{<1} \gamma(2) + \cdots + \underbrace{\left(1 - \frac{n-1}{n}\right)}_{<1} \gamma(n-1) \\
 &\quad + \underbrace{\left(1 - \frac{n-1}{n}\right)}_{<1} \gamma(-(n-1))
 \end{aligned}$$

por propiedad de valor absoluto y dado que cada $\left(1 - \frac{n-1}{n}\right) < 1$, entonces

$$\begin{aligned}
 &\leq |\gamma(0)| + |\gamma(1)| + |\gamma(-1)| + |\gamma(2)| + |\gamma(-2)| + \cdots + |\gamma(n-1)| + |\gamma(-(n-1))| \\
 &= \sum_{|h| < n} |\gamma(h)| = \sum_{-(n-1)}^{n-1} |\gamma(h)|
 \end{aligned}$$

con lo cual

$$nVar(\bar{X}_n) \leq \sum_{|h| < n} |\gamma(h)| \text{ implica que } Var(\bar{X}_n) \leq \frac{1}{n} \sum_{|h| < n} |\gamma(h)|$$

$$\begin{aligned}
 \frac{1}{n} \sum_{|h| < n} |\gamma(h)| &= \frac{1}{n} [|\gamma(0)| + |\gamma(-1)| + |\gamma(1)| + \cdots + |\gamma(n-1)| + |\gamma(-(n-1))|] \\
 &= \frac{1}{n} [\gamma(0) + 2|\gamma(1)| + 2|\gamma(2)| + \cdots + |\gamma(n-1)|] \tag{4.2}
 \end{aligned}$$

$$\begin{aligned}
 &= \frac{2}{n} [\underbrace{2\gamma(0)}_{a_1} + \underbrace{|\gamma(1)|}_{a_2} + \underbrace{|\gamma(2)|}_{a_3} + \cdots + \underbrace{|\gamma(n-1)|}_{a_n}] \tag{4.3}
 \end{aligned}$$

$$= \frac{2}{n} \sum_{j=1}^n a_j = 2 \frac{1}{n} \sum_{j=1}^n a_j$$

estableciendo que $b_n = n$ entonces $b_n \rightarrow \infty$ si $n \rightarrow \infty$. Ahora, como $a_j \rightarrow 0$ cuando $j \rightarrow \infty$ por hipótesis, usando el lema de Césaro tenemos que

$$\frac{1}{b_n} \sum_{k=1}^n (b_k - b_{k-1}) x_k \rightarrow x$$

si $b_n \rightarrow \infty$ y siempre que $x_n \rightarrow x$, cuando $n \rightarrow \infty$.

En particular como $x_n = |\gamma(n-1)| \rightarrow 0$, entonces

$$\frac{1}{n} \sum_{j=1}^n x_j \rightarrow x \quad x_n \rightarrow x,$$

así, $2 \lim_{n \rightarrow \infty} \left(\frac{1}{n} \sum_{j=1}^n a_j \right) = 0$

con lo cual $\text{var}(\bar{X}_n) \rightarrow 0$.

La segunda parte usando el lema de Kronecker, si $\{a_j\}$ es absolutamente sumable, es decir, $\lim_{n \rightarrow \infty} \sum_{j=-n}^n |a_j| < \infty$ entonces $\lim_{n \rightarrow \infty} \sum_{j=-n}^n \frac{j}{n} |a_j| = 0$

Así

$$\lim_{n \rightarrow \infty} n \text{Var}(\bar{X}_n) = \lim_{n \rightarrow \infty} \sum_{|h| < n} \left(1 - \frac{|h|}{n}\right) \gamma(h) \quad (4.4)$$

$$\begin{aligned} &= \lim_{n \rightarrow \infty} \sum_{|h| < n} \gamma(h) - \underbrace{\lim_{h \rightarrow \infty} \sum_{|h| < n} \frac{|h|}{n} \gamma(h)}_0 \\ &= \sum_{h=-\infty}^{\infty} \gamma(h). \end{aligned} \quad (4.5)$$

□

Nota 4.2. Si $X_t = \mu + \sum_{j=-\infty}^{\infty} \psi_j Z_{t-j}$ con $\sum_{j=-\infty}^{\infty} |\psi_j| < \infty$, entonces $\sum_{j=-\infty}^{\infty} |\gamma(h)| < \infty$ y así

$$n \text{Var}(\bar{X}_n) \rightarrow \sum_{h=-\infty}^{\infty} \gamma(h) = \sigma^2 \left(\sum_{j=-\infty}^{\infty} \psi_j \right)^2.$$

Nota 4.3. Lo que se ha probado básicamente es que \bar{X}_n converge en media cuadrática a μ y así en probabilidad al mismo límite. Asimismo, dado que para procesos ARMA(p, q) se satisfacen las condiciones del teorema anterior, entonces es posible ver que se cumple el teorema del límite central para proceso estocásticos estacionarios con la condición que $\sum_{j=-\infty}^{\infty} |\gamma(h)| < \infty$, y no sólo para procesos que son I.I.D.

Antes del teorema veamos la siguiente definición:

Definición 4.4. Una sucesión de VA's $\{X_n\}$ se dice que es asintóticamente normal con "media" μ_n y "desviación estándar" σ_n , si $\sigma_n > 0$, para n suficientemente grande y

$$\sigma_n^{-1}(X_n - \mu_n) \Rightarrow Z \quad Z \sim N(0, 1).$$

No es cierto que en general $\mu_n = E(X_n)$ o $\sigma_n^2 = \text{var}(X_n)$

Teorema 4.5. Si $\{X_t\}$ es un proceso estocástico estacionario tal que $X_t = \mu + \sum_{j=-\infty}^{\infty} \psi_j Z_{t-j}$ donde $\{Z_t\} \sim \text{IID}(0, \sigma^2)$ y $\sum_{j=-\infty}^{\infty} |\psi_j| < \infty$ y $\sum_{j=-\infty}^{\infty} \psi_j \neq 0$ entonces \bar{X}_n es $AN(\mu, n^{-1}v)$ con $v = \sum_{h=-\infty}^{\infty} \gamma(h) = \sigma^2 \left(\sum_{j=-\infty}^{\infty} \psi_j \right)^2$

4.2. ESTIMACIÓN DE LA FUNCIÓN DE AUTOCOVARIANZA Y DE AUTOCORRELACIÓN

Nota 4.6. Del teorema anterior tenemos que:

- ✓ Con este teorema uno puede encontrar intervalos de confianza aproximados para μ cuando el tamaño de la muestra es suficientemente grande.
- ✓ Por supuesto si $\{X_n\}$ es estacionario y Gaussiano tenemos que la distribución de \bar{X}_n es exacta y del segundo resultado del teorema 4.1 tenemos que

$$n^{-1/2}(\bar{X}_n - \mu) \sim N\left(0, \sum_{|h|<n} (1 - \frac{|h|}{n})\gamma(h)\right).$$

Sin embargo, si no se conoce a $\gamma(h)$, se puede cambiar una estimación $\hat{\gamma}(h)$, en vez de incluir todos los rezagos hasta n en la suma se restringe a \sqrt{n} o $n/4$ (Por Qué?).

- ✓ Si el proceso $\{X_t\}$ no es Gaussiano, entonces la distribución de \bar{X}_n es aproximada y se usa la misma expresión anterior.

Ejercicio 4.7. Lleve a cabo un estudio de simulación para verificar la que la confianza empírica de los intervalos de confianza para la media μ en un proceso ARMA(p,q) son aproximadamente del $100(1-\alpha)\%$. Qué sucede con la significancia si se omitiera la estructura de autocorrelación?

4.2 Estimación de la función de autocovarianza y de autocorrelación

Vamos a considerar los siguientes estimadores de $\gamma(h)$ y $\rho(h)$

$$\hat{\gamma}(h) = \frac{1}{n} \sum_{t=1}^{n-h} (X_t - \bar{X}_n)(X_{t+h} - \bar{X}_n) \quad 0 \leq h \leq n-1$$

y

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

$\hat{\gamma}(h)$ es un estimador sesgado de $\gamma(h)$. Note que para $h = 0$, el estimador es

$$\hat{\gamma}(0) = \frac{1}{n} \sum_{t=1}^n (X_t - \bar{X}_n)^2,$$

el cual tiene denominador n y no $n-1$ como es usual. Dividir entre $n-1$ asegura que el estimador sea insesgado. Sin embargo desde el punto de vista

asintótico ($n \rightarrow \infty$) la distribución de $\hat{\gamma}(h) - \gamma(h)$ tiene media cero, como lo veremos mas adelante.

Se puede verificar que para cada $n \geq 1$ la matriz

$$\hat{\Gamma}_n = \begin{bmatrix} \hat{\gamma}(0) & \hat{\gamma}(1) & \cdots & \hat{\gamma}(n-1) \\ \hat{\gamma}(1) & \hat{\gamma}(0) & \cdots & \hat{\gamma}(n-2) \\ \vdots & \vdots & \cdots & \vdots \\ \hat{\gamma}(n-1) & \hat{\gamma}(n-2) & \cdots & \hat{\gamma}(0) \end{bmatrix}$$

es definida no negativa ya que se puede verificar que

$$\hat{\Gamma}_n = n^{-1}TT'$$

donde T es una matriz $n \times 2n$

$$T = \begin{bmatrix} 0 & \cdots & 0 & Y_1 & Y_2 & \cdots & Y_n \\ 0 & \cdots & 0 & Y_1 & Y_2 & \cdots & Y_n & 0 \\ \vdots & & & & & & \vdots \\ 0 & Y_1 & Y_2 & \cdots & Y_n & 0 & \cdots & 0 \end{bmatrix}$$

con $Y_i = X_i - \bar{X}_n$ $i = 1, 2, 3, \dots, n$. Luego, para cada vector $\underbrace{a}_{\sim} \in \mathbb{R}^n$ se puede verificar que

$$\underbrace{a'}_{\sim} \hat{\Gamma}_n \underbrace{a}_{\sim} = n^{-1}(\underbrace{a'}_{\sim} T)(\underbrace{a'}_{\sim} T)' \geq 0.$$

También la matriz de correlación

$$\hat{R}_n = \frac{\hat{\Gamma}_n}{\hat{\gamma}(0)}$$

es definida no negativa.

Nota 4.8. En ocasiones, en la definición de $\hat{\gamma}(h)$, es reemplazado el factor n^{-1} por $(n-h)^{-1}$, pero las matrices $\hat{\Gamma}_n$ y \hat{R}_n pueden no ser definidas no-negativas.

Nota 4.9. Se puede verificar que $\hat{\Gamma}_n > 0$ si $\hat{\gamma}(0) > 0$

Nota 4.10. Desde el punto de vista práctico tenemos lo siguiente:

- Basado en solo la información X_1, X_2, \dots, X_n , no es posible estimar $\gamma(k)$ para $k \geq n$.
- Para valores de k un poco mas pequeños que n los estimadores $\hat{\gamma}(k)$ no son muy confiables ya que de sólo intervienen un poco de pares (X_t, X_{t+k}) , por ejemplo para $k = n-1$ solo se usa un único par.

4.2. ESTIMACIÓN DE LA FUNCIÓN DE AUTOCOVARIANZA Y DE AUTOCORRELACION

- Box-Jenkins sugieren que son útiles las estimaciones de la autocorrelación $\rho(k)$ puede ser hechas cuando n es mayor a 50 y para valores de $k \leq \frac{n}{4}$ o $k \leq \sqrt{n}$.

Una de las cosas importantes al identificar un modelo ARMA(p,q) para un conjunto de observaciones, es que se identifiquen cuales autocorrelaciones poblacionales(basadas en las muestrales) son significativamente diferentes de 0. Para eso es útil el siguiente resultado:

Teorema 4.11. Si $\{X_t\}$ es un proceso estocástico estacionario tal que

$$X_t - \mu = \sum_{j=-\infty}^{\infty} \psi_j Z_{t-j} \quad \{Z_t\} \sim IID(0, \sigma^2)$$

donde $\sum_{j=-\infty}^{\infty} |\psi_j| < \infty$ y $EZ_t^4 < \infty$, entonces para cada $h = 1, 2, 3, \dots$ tenemos que

$$\hat{\rho}(h) \text{ es } AN(\rho(h), n^{-1}W)$$

donde

$$\hat{\rho}(h) = (\hat{\rho}(1), \hat{\rho}(2), \dots, \hat{\rho}(h))$$

$$\rho(h) = (\rho(1), \rho(2), \dots, \rho(h))$$

y W es la matriz de covarianza, tiene entradas

$$\begin{aligned} W_{ij} = & \sum_{k=-\infty}^{\infty} \{ \rho(k+i)\rho(k+j) + \rho(k-i)\rho(k+j) + 2\rho(i)\rho(j)\rho^2(k) \\ & - 2\rho(i)\rho(k)\rho(k+j) - 2\rho(j)\rho(k)\rho(k+i) \} \end{aligned}$$

Una forma de relajar el supuesto que $E[Z_t^4] < \infty$ es dada en el siguiente teorema.

Teorema 4.12. Si $\{X_t\}$ es estacionario tal que

$$X_t - \mu = \sum_{j=-\infty}^{\infty} \psi_j Z_{t-j} \quad \{Z_t\} \sim RB(0, \sigma^2)$$

donde $\sum_{j=-\infty}^{\infty} |\psi_j| < \infty$ y $\sum_{j=-\infty}^{\infty} \psi_j^2 |j| < \infty$, entonces para cada $h \in \{1, 2, \dots\}$

$$\hat{\rho}(h) \text{ es } AN(\rho(h), n^{-1}W)$$

Nota 4.13. (*Fórmula de Bartlett*)

Se puede verificar que

$$W_{ij} = \sum_{k=1}^{\infty} \{\rho(k+i) + \rho(k-i) - 2\rho(i)\rho(k)\} \times \{\rho(k+j) + \rho(k-j) - 2\rho(j)\rho(k)\} \quad (4.6)$$

Nota 4.14. Se puede verificar las condiciones del teorema anterior son satisfechas para un modelo ARMA(p,q) con proceso $\{Z_t\}$ IID de media cero y varianza finita.

Nota 4.15. La prueba del Teorema 4.11 se hace primero, verificando que la distribución asintótica del vector aleatorio $[\hat{\gamma}^*(0), \dots, \hat{\gamma}^*(h)]'$ conformado por

$$\hat{\gamma}^*(h) = \frac{1}{n} \sum_{t=1}^n X_t X_{t+h}, \quad h = 0, 1, 2, \dots$$

es normal multivariada usando la proposición 6.3.9 del libro [3]. Luego se demuestra que los vectores aleatorios $[\hat{\gamma}^*(0), \dots, \hat{\gamma}^*(h)]'$ y $[\hat{\gamma}(0), \dots, \hat{\gamma}(h)]'$ tienen la misma distribución asintótica. Finalmente se usa proposición 6.4.3 del libro [3] para probar que la transformación de $[\hat{\gamma}(0), \dots, \hat{\gamma}(h)]'$ dada por $g([x_0, x_1, \dots, x_h]') = [x_1/x_0, \dots, x_h/x_0]$ también es asintóticamente normal multivariada.

Ejemplo 4.16. Si $\{X_t\} \sim \text{IID}(0, \sigma^2)$, entonces $\rho(l) = 0 \quad \text{si} \quad |l| > 0$.

Así

$$W_{ij} = \begin{cases} 1 & \text{si } i = j \\ 0 & \text{e.o.c} \end{cases}$$

ya que $\rho(k+i)$ y $\rho(i)\rho(k)$ son cero para todo k y todo i , mientras que $\rho(k-i)$ y $\rho(k-j)$ es $\neq 0$ si $k = j = i$

Para n suficientemente grande $\hat{\rho}(1), \dots, \hat{\rho}(h)$ son variables aleatorias aproximadamente independientes e idénticamente distribuidas normales de media cero y varianza n^{-1} . Así un IC aproximado para cada $\rho(1), \dots, \rho(h)$ es

$$\hat{\rho}(h) \pm Z_{1-\alpha/2} n^{-1/2}.$$

El anterior procedimiento puede ser usado para comprobar si las observaciones en verdad provienen de un proceso IID, o más específicamente si es no autocorrelacionado hasta el rezago h , es decir $\rho(1) = \rho(2) = \dots = \rho(h) = 0$. La siguiente gráfica muestra un ejemplo:

4.2. ESTIMACIÓN DE LA FUNCIÓN DE AUTOCOVARIANZA Y DE AUTOCORRELACION

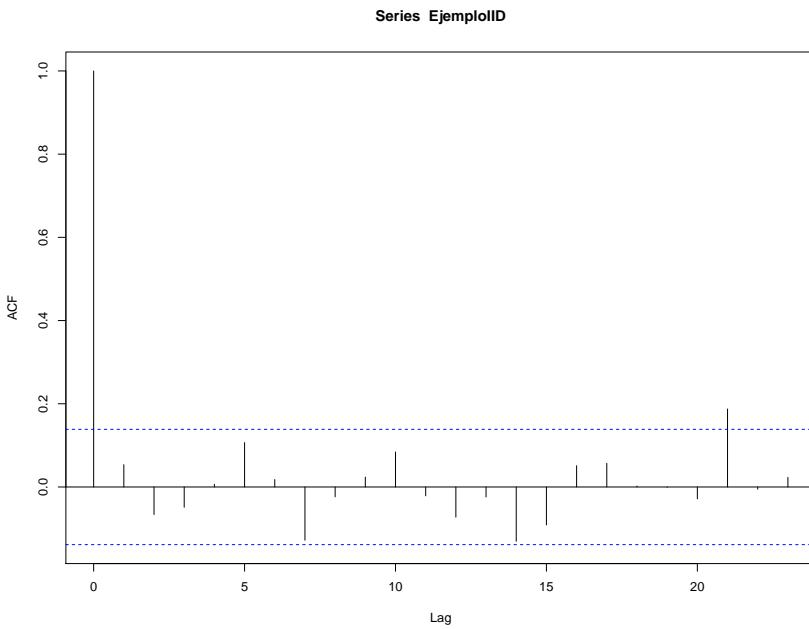


Figura 4.1: Gráfica ACF de una serie IID en R usando acf

Ejemplo 4.17. Consideremos que $\{X_t\}$ es un proceso MA(1) tal que

$$X_t = Z_t + \theta Z_{t-1}.$$

Usando la fórmula en 4.6, se puede verificar que las entradas de la diagonal principal de la matriz W quedan determinadas como sigue:

$$w_{ii} = \begin{cases} 1 - 3\rho^2(1) + 4\rho^4(1), & \text{si } i = 1, \\ 1 + 2\rho^2(1), & \text{si } i > 1, \end{cases}$$

lo que quiere decir, es que estas son las varianzas aproximadas de $n^{-1/2}(\hat{\rho}(i) - \rho(i))$. Para ver si los datos de la muestra X_1, \dots, X_n son compatibles con un MA(1), primero verificamos que no sea IID, usando la metodología anterior. Después, como se sabe que en un MA(1) $\rho(1) \neq 0$ y $\rho(i) = 0$ para $i > 1$, entonces procedemos a probar que la autocorrelación $\rho(1)$ sea significativa, mientras que las autocorrelaciones para rezagos más altos no lo sean. Por lo tanto, usaremos las varianzas asintóticas para probar su significancia. Para probar la hipótesis nula $H_0 : \rho(1) = 0$, entonces calculamos el estadístico de

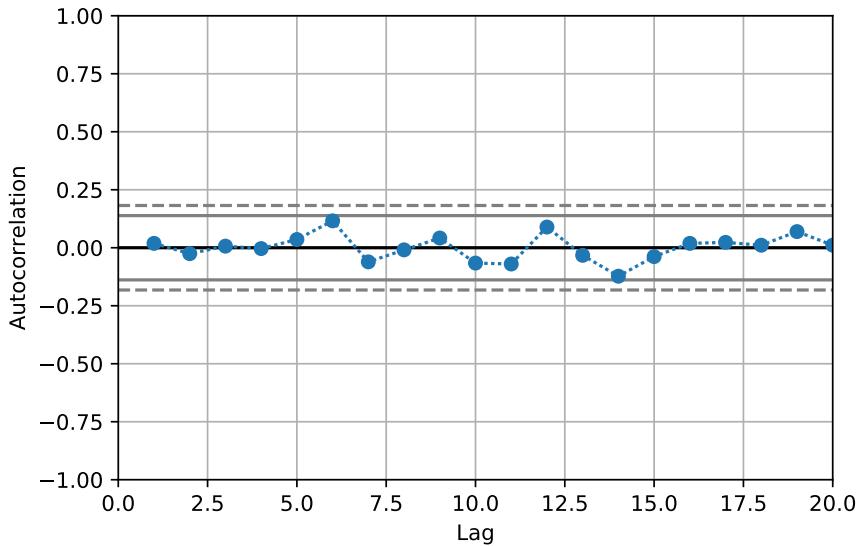


Figura 4.2: Gráfica ACF de una serie IID en Python Usando autocorrelation plot de Pandas

prueba bajo la hipótesis nula

$$Z = \frac{\hat{\rho}(1) - 0}{\frac{1}{n^{1/2}} \sqrt{1 - 3\rho^2(1) + 4\rho^4(1)}} \quad \text{es} \quad AN(0, 1),$$

con lo cual se debería rechazar H_0 si $|z| > z_{1-\alpha/2}$. Como los valores de las correlaciones en el denominador son desconocidos para computar la varianza, entonces usamos sus estimaciones. Lo cual es equivalente a probar que

$$|\hat{\rho}(1)| > z_{1-\alpha/2} \frac{1}{n^{1/2}} \sqrt{1 - 3\hat{\rho}^2(1) + 4\hat{\rho}^4(1)}.$$

Lo que quiere decir, es que si $\hat{\rho}(1)$ está por fuera de las bandas de confianza $\pm z_{1-\alpha/2} \frac{1}{n^{1/2}} \sqrt{1 - 3\hat{\rho}^2(1) + 4\hat{\rho}^4(1)}$ entonces es significativamente diferente de cero. Con un razonamiento igual se puede verificar la significancia de $\rho(i)$ para $i \geq 2$ usando unas bandas $\pm z_{1-\alpha/2} \frac{1}{n^{1/2}} \sqrt{1 + 2\hat{\rho}^2(1)}$.

Veamos un ejemplo en R.

La gráfica 4.3 nos muestra el ACF simple de bandas consistente con un proceso MA, para una serie simulada que proviene de un proceso MA(1). Note que la gráfica es consistente para identificar el proceso MA(1). La gráfica 4.4 es una gráfica obtenida usando Python.

4.2. ESTIMACIÓN DE LA FUNCIÓN DE AUTOCOVARIANZA Y DE AUTOCORRELACION

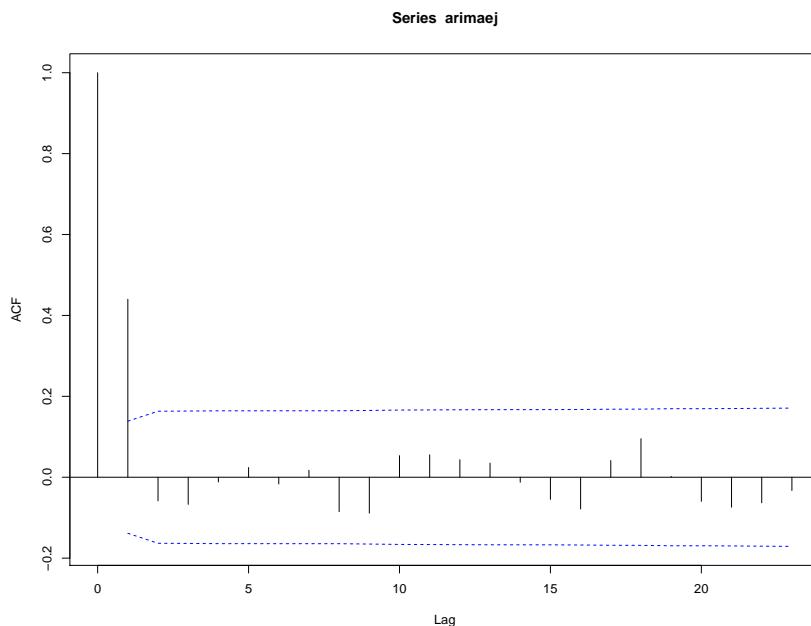


Figura 4.3: ACF en R con bandas de confianza para un MA(1)

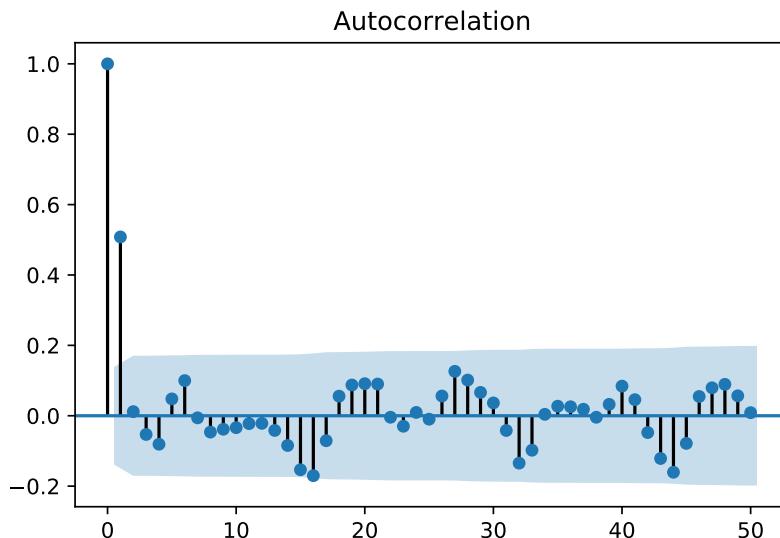


Figura 4.4: ACF en Python con bandas de confianza para un MA(1)

Nota 4.18. Si

$$X_t = Z_t + \theta_1 Z_{t-1} + \cdots + \theta_q Z_{t-q} \quad \{Z_t\} \sim IID(0, \sigma^2)$$

se puede verificar de la formula de Barlett que

$$w_{ii} = [1 + 2\rho^2(1) + 2\rho^2(2) + \cdots + 2\rho^2(q)] \quad i > q$$

por lo cual la varianza asintótica de $n^{-1/2}(\hat{\rho}(i) - \rho(i))$ n es suf. grande. Lo anterior permite diseñar un procedimiento para identificar un proceso MA(q) usando la función de autocorrelación simple.

Veamos que éste mismo procedimiento no es adecuando para el caso de procesos autoregresivos.

Ejemplo 4.19. Podemos aplicar la formula de Barlett 4.6 a un proceso causable AR(1)

$$X_t - \phi X_{t-1} = Z_t \quad \{Z_t\} \sim IID(0, \sigma^2).$$

Cómo se sabe que $\rho(i) = \phi^{|i|}$, entonces la varianza asintótica de $n^{-1/2}(\hat{\rho}(i) - \rho(i))$ es

$$w_{ii} = \sum_{k=1}^i \phi^{2k} (\phi^{-k} - \phi^k)^2 + \sum_{k=i+1}^{\infty} \phi^{2k} (\phi^i - \phi^k)^2 \quad (4.7)$$

$$= (1 - 2\phi^{2i})(1 + \phi^2)(1 - \phi^2)^{-1} - 2i\phi^{2i} \quad (4.8)$$

$$\simeq (1 + \phi^2)/(1 - \phi^2) \quad \text{para } i \text{ grande}$$

No podemos sugerir una metodología igual que antes por que para ningún i $\rho(i) = 0$, lo cual hace que no tenga sentido proponer bandas para probar $\rho(i) = 0$. Esto se hará usando la función de autocorrelación parcial FACP.

Ejemplo 4.20. Del ejemplo 3.40 tenemos que para un proceso AR(p) puro, la función autocorrelación parcial(FACP) teórica satisface que $\alpha(k) = 0$ para $k > p$. Por otro lado, como se definió que la FACP muestral $\hat{\alpha}(k) = \hat{\phi}_{kk}$, se puede verificar usando Teorema 8.1.2 del libro [3] que

$$n^{1/2} \hat{\phi}_{mm} \Rightarrow N(0, 1)$$

para $m > p$ en un proceso AR(p) y $n \rightarrow \infty$. Con lo cual, si un modelo autoregresivo es apropiado para los datos, entonces para un rezago finito los valores observador $\hat{\phi}_{mm}$ deberían ser compatibles con la distribución $N(0, 1/n)$. En particular, si el orden del proceso es p , entonces para $m > p$, $\hat{\phi}_{mm} = \hat{\alpha}(m)$ caerán dentro de las bandas $\pm z_{1-\alpha/2} n^{-1/2}$, mientras que al menos para $m=p$, $\hat{\alpha}(m)$ debería estar por fuera de estas bandas.

4.2. ESTIMACIÓN DE LA FUNCIÓN DE AUTOCOVARIANZA Y DE AUTOCORRELACION

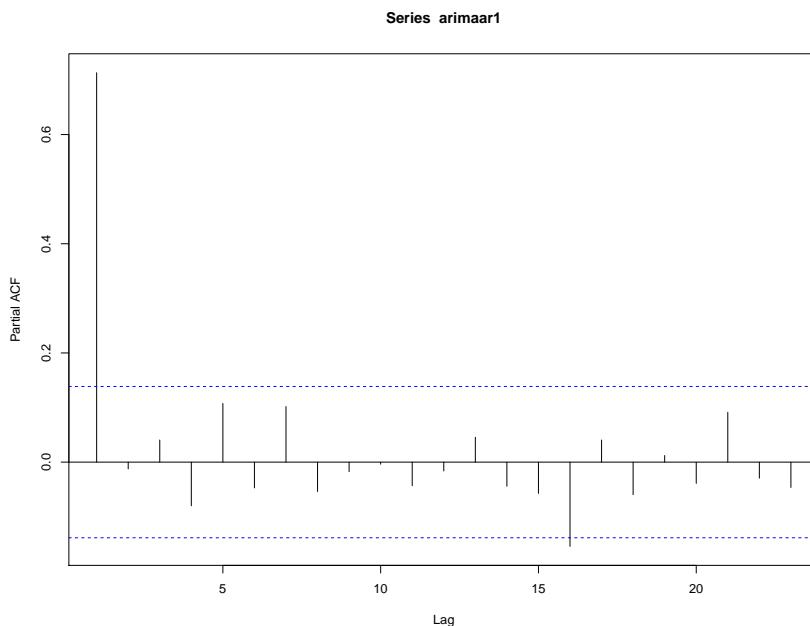


Figura 4.5: PACF de una serie simulada de un proceso AR(1) en R

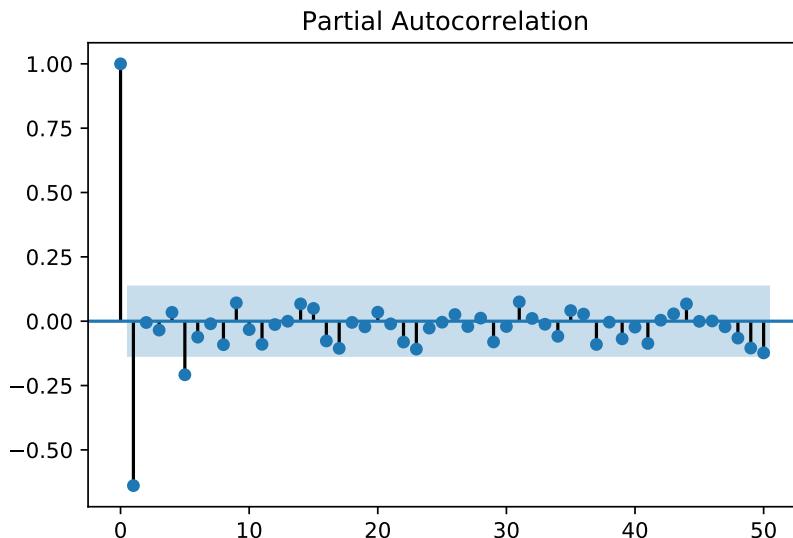


Figura 4.6: PACF de una serie simulada de un proceso AR(1) en Python

Veamos un ejemplo simulado para ver como funciona la función de autocorrelación parcial para identificar si un proceso proviene de un AR puro. En la figura 4.5 se puede ver la PACF del ejemplo simulado. Note que es consistente con un AR(1). Una gráfica similar desarrollada en Python puede verse en 4.6. Usaremos los archivos de R y Phyton con el nombre IdentificacionMAyARpueros.R o IdentificacionMAyARpueros.ipynb.

Ejercicio 4.21. *Que sucede si usted hace el ACF de un $AR(p)$? Qué sucede si usted hace el PACF de un $MA(q)$? Lleve a cabo simulaciones para ver éste hecho.*

5

Estimación de Procesos ARMA

Por ahora sólo hemos discutido que nuestra series $\{x_1, \dots, x_T\}$ proviene de un proceso estocástico estacionario. Sin embargo en general eso no es cierto. Por lo pronto haremos la estimación de los parámetros de un modelo ARMA(p,q), los cuales permiten modelar una serie de tiempo estacionaria. Al final del capítulo, introduciremos el modelamiento de series de tiempo no estacionarias vía modelos diferentes a la extensión de modelos ARMA para procesos no estacionarios.

5.1 Estimación Vía Máxima Verosimilitud

Existe varios métodos para la estimación de los parámetros de un modelo ARMA(p,q), por ejemplo a través de las ecuaciones de Yule-Walker, del algoritmo Durbin-Levinson o inclusive vía mínimos cuadrados. Sin embargo vamos a enfocarnos en el método de máxima verosimilitud ya que ofrece buenas propiedades asintóticas y es el que usualmente implementan los programas de computadora.

Recordemos que la función de verosimilitud se define a través de la función de distribución conjunta del vector (X_1, \dots, X_n) , es decir,

$$L(\phi, \theta | X_1, \dots, X_n, p, q) = f_{X_1, \dots, X_n}(x_1, \dots, x_n; \phi, \theta | p, q) \quad (5.1)$$

En forma general, para procesos Gaussianos estacionarios tenemos que la función de verosimilitud del vector (X_1, \dots, X_n) puede escribirse como

$$L(\Gamma_n) = (2\pi)^{-n/2} \det(\Gamma_n)^{-1/2} \exp\left(-\frac{1}{2} X_n' \Gamma_n^{-1} X_n\right) \quad (5.2)$$

donde $\Gamma_n = E[X_n, X_n']$ es no singular. La verosimilitud 5.2 se puede escribir equivalente como

$$L(\Gamma_n) = (2\pi)^{-n/2} (\nu_0 \nu_1 \dots \nu_{n-1})^{-1/2} \exp\left(-\frac{1}{2} \sum_{j=1}^n (X_j - \hat{X}_j)^2 / \nu_{j-1}\right) \quad (5.3)$$

donde ν_j son los errores cuadráticos medios de los predictores un paso adelante \hat{X}_j de X_j .

5.2 Función de Verosimilitud de un Proceso ARMA(p,q)

Sea $\{X_t\}$ un proceso estocástico estacionario que sigue un modelo ARMA(p,q) causal tal que

$$X_t - \phi_1 X_{t-1} - \dots - \phi_p X_{t-p} = Z_t + \theta_1 Z_{t-1} + \dots + \theta_q Z_{t-q}$$

tal que $\{Z_t\} \sim RB(0, \sigma^2)$.

Los parámetros a estimar son las ordenes p y q, los vectores $\phi = (\phi_1, \dots, \phi_p)'$, $\theta = (\theta_1, \dots, \theta_q)'$ y σ^2 . Vamos a asumir que el proceso $\{Z_t\}$ es Gaussiano, lo cual implica que el proceso $\{X_t\}$ sea Gaussiano por el supuesto de causalidad. Si un modelo con media diferente de cero es especificado, una constante c debe ser incluida en el modelo y por lo tanto también necesita ser estimada. Recordar que basados en esa constante c , se tiene que $E[X_t] = \frac{c}{1-\phi_1-\phi_2-\dots-\phi_p}$. En principio asumiremos que los órdenes p, q son conocidos, y entonces no enfocaremos en los otros parámetros para su estimación.

Vamos a asumir que se tiene disponible una muestra $\{X_1, \dots, X_n\}$ de un proceso ARMA(p,q) Guassiano de media cero(sin pérdida de generalidad). Anteriormente se mostró que los predictores un paso adelante \hat{X}_{i+1} para modelos ARMA(p,q) son dados por

$$\hat{X}_{i+1} = \sum_{j=1}^i \theta_{ij} (X_{i+1-j} - \hat{X}_{i+1-j}) \quad 1 \leq i < m = \max(p, q)$$

$$\hat{X}_{i+1} = \phi_1 X_i + \dots + \phi_p X_{i-p} + \sum_{j=1}^q \theta_{ij} (X_{i+1-j} - \hat{X}_{i+1-j}) \quad i \geq m$$

Y los errores cuadráticos medios son:

$$E[(X_j - \hat{X}_j)^2] = \sigma^2 r_i \quad r_i = \frac{\nu_i}{\sigma^2}$$

Nota 5.1. Los θ_{ij} y r_i no depende de σ^2 ya que las covarianzas del proceso $\{W_t\}$ en 3.27 depende sólo de ϕ_1, \dots, ϕ_p y $\theta_1, \dots, \theta_q$ con lo cuál \hat{X}_j y r_j no depende de σ^2 . Se involucran en el cálculo pero no afecta la magnitud de la cantidad con lo cuál, la verosimilitud Gaussiana del vector $X_n = (X_1, \dots, X_n)'$ es

$$L(\phi, \theta, \sigma^2) = (2\pi\sigma^2)^{-n/2} (r_0, \dots, r_{n-1})^{-1/2} \exp \left[-\frac{1}{2\sigma^2} \sum_{j=1}^n (X_j - \hat{X}_j)^2 / r_j \right] \quad (5.4)$$

Se puede verificar que diferenciando parcialmente a $\ln(L(\phi, \theta, \sigma^2))$ con respecto a σ^2 , y dada la independencia de \hat{X}_j y r_j con σ^2 , los estimadores de máxima verosimilitud satisfacen

$$\hat{\sigma}^2 = n^{-1} S(\hat{\phi}, \hat{\theta}) \quad \text{donde } S(\hat{\phi}, \hat{\theta}) = \sum_{j=1}^n (X_j - \hat{X}_j)^2 / r_j$$

y $\hat{\phi}, \hat{\theta}$ son los valores de ϕ, θ que minimizan

$$l(\phi, \theta) = \ln(n^{-1} S(\hat{\phi}, \hat{\theta})) + n^{-1} \sum_{j=1}^n \ln r_{j-1}.$$

A $l(\phi, \theta)$ se llama la verosimilitud reducida. La idea es finalmente encontrar los estimadores de ϕ, θ que minimizan $l(\phi, \theta)$ usando un algoritmo iterativo de optimización junto con el algoritmo de innovaciones, y luego reemplazar en $n^{-1}(S(\hat{\phi}, \hat{\theta}))$ para obtener el estimador de σ^2 .

Nota 5.2. La mayoría de paquetes de computadora usan la representación en modelos de espacio estado de un modelo ARMA(p,q), para luego escribir la verosimilitud Gaussiana en términos del modelos de espacio y estado. Esta escritura en modelos de espacio estado permite obtener de una forma mas eficiente las predicciones un paso adelante y también la optimización de la función de verosimilitud. Mas adelante veremos los modelos de espacio estado.

5.3 Propiedades Asintóticas de los estimadores de máxima verosimilitud

Si $\{X_t\}$ es un proceso causal e invertible tal que

$$X_t = \phi_1 X_{t-1} - \cdots - \phi_p X_{t-p} = Z_t + \theta_1 Z_{t-1} + \cdots + \theta_q Z_{t-q} \quad \{Z_t\} \sim IID(0, \sigma^2),$$

con $\phi(\cdot)$ y $\theta(\cdot)$ sin ceros comunes, entonces el estimador de máxima verosimilitud $\hat{\beta} = (\hat{\phi}_1, \dots, \hat{\phi}_p, \hat{\theta}_1, \dots, \hat{\theta}_q) = (\hat{\phi}, \hat{\theta})$ es aquel valor de $\beta = (\phi, \theta)$ que maximiza la verosimilitud reducida $l(\phi, \theta)$.

Por supuesto en la práctica los valores puntuales son obtenidos de forma numérica.

Nota 5.3. Los estimadores de mínimos cuadrados ϕ, θ son aquellos que minimizan la $\ln(n^{-1}s(\phi, \theta)) = l(\phi, \theta) - n^{-1} \sum_{j=1}^n \ln r_{j-1}$.

El término $n^{-1} \sum_{j=1}^n \ln r_{j-1}$ es insignificante cuando $n \rightarrow \infty$, lo cual hace que los estimadores de mínimos cuadrados y máxima verosimilitud tengan propiedades asintóticas iguales. Bajo estas condiciones se puede verificar que

$$n^{1/2}(\hat{\beta} - \beta) \Rightarrow N(0, V(\beta))$$

Nota 5.4. Para obtener el resultado anterior hay que usar resultados de la teoría espectral.

Nota 5.5. Si $p \geq 1$ y $q \geq 1$

$$V(\beta) = \sigma^2 \begin{bmatrix} E[U_t U'_t] & E[U_t V'_t] \\ E[v_t U'_t] & E[v_t V'_t] \end{bmatrix}^{-1}$$

donde $u_t = (u_t, \dots, u_{t+1-p})'$, $v_t = (v_t, \dots, v_{t+1-q})'$ donde $\{u_t\}$ y $\{v_t\}$ son procesos autoregresivos no correlacionados tal que

$$\phi(B)V_t = Z_t \quad y \quad \theta(B)V_t = Z_t$$

para $p=0$ $v(\beta) = \sigma^2 [E(V_t V'_t)]^{-1}$ y para $q=0$ $v(\beta) = [\sigma^2 [E(U_t, U'_t)]^{-1}]$.

Salvo en algunos casos particulares, expresiones analíticas exactas puede ser encontradas de los estimadores de máxima verosimilitud.

Ejemplo 5.6. Estimación en un proceso AR(1)

Para el caso $AR(1)$ se puede encontrar de manera explícita la forma que tienen los estimadores de máxima verosimilitud el cual se basa en la muestra

5.3. PROPIEDADES ASINTÓTICAS DE LOS ESTIMADORES DE MÁXIMA VERSIÓN II

Y_1, \dots, Y_T .

Sea $Y_t = c + \phi Y_{t-1} + \epsilon_t$ con $|\phi| < 1$ y $\{\epsilon_t\} \sim IIDN(0, \sigma^2)$. Por lo tanto el vector de parámetro queda establecido como

$$\theta = (c, \phi, \sigma^2).$$

Ahora, como $\{Y_t\}$ es causable entonces $\{Y_t\}$ es Gaussiano por ser $\{\epsilon_t\}$ Gaussiano, con lo cual

$$E[Y_1] = \mu = E[c + \phi Y_0 + \epsilon_1] = c + \phi E[Y_0] + 0$$

con lo cual

$$\mu = c + \phi\mu \quad (5.5)$$

$$\mu - \phi\mu = c \quad (5.6)$$

$$\mu(1 - \phi) = c \quad (5.7)$$

$$\mu = \frac{c}{1 - \phi}.$$

También, se puede ver que $V[Y_1] = \frac{\sigma^2}{1 - \phi^2} = \gamma(0)$ es decir,

$$f_{Y_1, \phi}(y_1) = \frac{1}{\sqrt{2\pi}\sqrt{\sigma^2/(1 - \phi^2)}} \exp\left[-\frac{1}{2} \frac{\{y_1 - [c/(1 - \phi)]\}^2}{\sigma^2/(1 - \phi^2)}\right]$$

La verosimilitud basada Y_1, \dots, Y_T es

$$f_{Y_1, \dots, Y_T}(y_1, y_2, \dots, y_T) =$$

$$f_{Y_1}(y_1) \cdot f_{Y_2|Y_1}(y_2|y_1) f_{Y_3|Y_2, Y_1}(Y_3|y_2, y_1) \cdots f_{Y_T|Y_{T-1}, \dots, Y_1}(y_T|y_{T-1}, \dots, y_1).$$

Se puede verificar que cada densidad satisface

$$f_{Y_t|Y_{t-1}, \dots, Y_1}(y_t|y_{t-1}, \dots, y_1; \theta) = f_{Y_t|Y_{t-1}}(y_t|y_{t-1}; \theta)$$

dado que el proceso $\{Y_t\}$ es autoregresivo. Note que

$$E[Y_t|Y_{t-1}] = E[c + \phi Y_{t-1} + \epsilon_t|Y_{t-1} = y_{t-1}] \quad (5.8)$$

$$= c + \phi Y_{t-1} + E[\epsilon_t] \quad (5.9)$$

$$= c + \phi Y_{t-1},$$

y que

$$Var(Y_t|Y_{t-1}) = Var(\epsilon_t) = \sigma^2.$$

Así

$$\begin{aligned} f_{Y_t|Y_{t-1}=y_{t-1}}(y_t|Y_{t-1}=y_{t-1}; \theta) &\underset{\text{corresponde}}{\approx} N(c + \phi y_{t-1}, \sigma^2) \\ &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left[-\frac{1}{2} \frac{(y_t - c - \phi y_{t-1})^2}{\sigma^2} \right] \end{aligned} \quad (5.10)$$

con lo cual

$$f_{Y_1, \dots, Y_T}(y_1, \dots, y_T) = f_{Y_1}(y_1; \theta) \prod_{t=2}^T f_{Y_t|Y_{t-1}}(y_t|y_{t-1}; \theta).$$

Ahora, el logaritmo de la función de verosimilitud es

$$L(\theta) = \log(f_{Y_1}(y_1; \theta)) + \sum_{t=2}^T \log f_{Y_t|Y_{t-1}}(y_t|y_{t-1}; \theta).$$

En ocasiones se prefiere la verosimilitud condicional Y_1 es decir

$$f_{Y_2, Y_3, \dots, Y_T}(y_2, \dots, y_T|y_1; \theta) = \prod_{t=2}^T f_{Y_{t-1}|Y_{t-1}}(y_{t-1}|y_{t-1}; \theta)$$

$$\begin{aligned} L^*(\theta) &= \sum_{t=2}^T \log(f_{Y_t|Y_{t-1}}(y_t|y_{t-1}; \theta)) \\ &= [(T-1)/2]\log(2\pi) - [(T-1)/2]\log(\sigma^2) - \sum_{t=2}^T \left[\frac{(y_t - c - \phi y_{t-1})^2}{2\sigma^2} \right]. (*) \end{aligned}$$

Maximizar (*) es equivalente a maximizar

$$\sum_{t=2}^T (y_t - c - \phi y_{t-1})^2 = (y - X\beta)'(y - X\beta)$$

Con respecto a c y ϕ el cual es claramente una regresión lineal ordinaria de y_t sobre una constante y su rezago, donde

$$y = \begin{bmatrix} y_2 \\ y_3 \\ \vdots \\ y_T \end{bmatrix} \quad X = \begin{bmatrix} 1 & y_1 \\ 1 & y_2 \\ \vdots & \\ 1 & y_{T-1} \end{bmatrix} \quad \beta = \begin{bmatrix} c \\ \phi \end{bmatrix}.$$

El estimador de máxima verosimilitud para c y ϕ son

$$\begin{bmatrix} \hat{c} \\ \hat{\phi} \end{bmatrix} = \begin{bmatrix} T-1 & \sum_{t=2}^T y_{t-1} \\ \sum_{t=2}^T y_{t-1} & \sum_{t=2}^T y_{t-1}^2 \end{bmatrix}^{-1} \begin{bmatrix} \sum_{t=2}^T y_{t-1} \\ \sum_{t=2}^T y_{t-1} y_t \end{bmatrix}$$

y el estimador de máxima verosimilitud para σ^2 es

$$\hat{\sigma}^2 = \sum \left[\frac{(y_t - \hat{c} - \hat{\phi}y_{t-1})^2}{T-1} \right] = \frac{\sum_{t=2}^T \hat{\epsilon}_t^2}{T-1},$$

donde $\hat{\epsilon}_t$ es el error de predicción un paso adelante intra-muestra.

5.4 Identificación de los ordenes p,q y análisis de residuales

Si en el estado de la naturaleza el proceso $\{X_t\}$ sigue un proceso ARMA(p,q) mixto, entonces no es razonable usar los criterios del capítulo anterior para identificar plenamente los órdenes p y q. Para esto existen algunas metodologías diferentes como lo son los criterios de información(AIC, AICc, BIC, etc), la complejidad y entropía , la función de autocorrelación extendida, validación cruzada, etc. En ésta sección hablaremos de los criterios de información como una herramienta para elegir modelos, la cual se usará en este caso para identificar los ordenes p y q.

5.4.1 Criterios de Información

En 1973, Akaike propuso un estimador aproximadamente insesgado del índice Kullback-Liebler del modelo ajustado relativo al verdadero modelo. Si $\tilde{\mathbf{X}}$ es un vector aleatorio cuya función de densidad pertenece a la familia $\{f(\cdot|\psi) : \psi \in \Psi\}$. La discrepancia **Kullback-Liebler** entre $f(\cdot|\psi)$ y $f(\cdot|\theta)$ se define como la divergencia Kullback-Liebler

$$d(\psi|\theta) = \Delta(\theta|\theta) - \Delta(\psi|\theta)$$

lo cual es equivalente a escribir

$$d(\psi|\theta) = \Delta(\theta|\theta) + \frac{1}{2}(-2\Delta(\psi|\theta))$$

donde

$$\Delta(\psi|\theta) = E_\theta[\ln f(\tilde{\mathbf{X}}; \psi)] = \int_{\mathbb{R}^n} \ln(f(\tilde{\mathbf{x}}; \psi)) f(\tilde{\mathbf{x}}; \theta) d\tilde{\mathbf{x}}$$

se conoce como el índice del Kullback-Liebler de $f(\cdot|\psi)$ relativo a $f(\cdot|\theta)$. Se puede verificar que $d(\psi|\theta) \geq 0$ con la igualdad si y sólo si $f(\tilde{\mathbf{x}}; \psi) = f(\tilde{\mathbf{x}}; \theta)$ lo cual implica que $\Delta(\psi|\theta) = \Delta(\theta|\theta)$.

Nota 5.7. • Dadas las observaciones X_1, \dots, X_n de un proceso ARMA con parámetros desconocidos $\theta = (\beta, \sigma)$, el verdadero modelo puede ser identificado si fuera posible computar la discrepancia Kullback-Liebler entre todos los modelos candidatos y el verdadero modelo.

- No es posible computar la discrepancia porque θ es desconocido. Sin embargo, no es necesario computar la discrepancia de forma exacta ya que si hacemos que el índice $\Delta(\psi|\theta)$ sea máximo o equivalentemente que sea mínimo $-2\Delta(\psi|\theta)$ debería estar muy cerca del índice $\Delta(\theta|\theta)$, lo cual hace que la discrepancia sea cercana a cero.
- Una idea de minimizar el índice, está relacionada con evaluar el índice del Kullback-Liebler en los estimadores que hacen máxima la función de verosimilitud o mínima el logaritmo de la función de verosimilitud (es decir, este proceso está relacionada con los estimadores de máxima verosimilitud).
- Si computamos $(-2\Delta(\psi|\theta))$ minimizado para cada modelo, entonces elegimos el modelo cuya cantidad se haga mínimo.
- Asumamos que el verdadero modelo y los alternativos son Gaussianos, y sea $\theta = (\beta, \sigma^2)$ con $f(\cdot|\theta)$ la densidad conjunta de $(X_1, \dots, X_n)'$ con $\{X_t\}$ un proceso ARMA(p, q) Gaussiano.
- Note que θ depende de p, q a través de su dimensión.
- La idea entonces de Akaike, consistió en encontrar un estimador inscrito de la cantidad $E[-2\Delta(\hat{\theta}|\theta)]$ (el índice de Kullback-Liebler esperado), donde $\hat{\theta}$ es el estimador de máxima verosimilitud de θ cuyos órdenes son p, q .
- Entonces podemos seleccionar los valores de p y q para nuestro modelo ajustado en donde se minimice el criterio de información AICC($\hat{\beta}$), donde

$$AICC(\beta) = -2 \ln L_X(\beta, S_X(\beta)/n) + \frac{2(p+q+1)n}{n-p-q-2}.$$

- De cierta manera el criterio de información termina siendo una medida para llevar a cabo validación cruzada como lo menciona [5] en el capítulo 6.

- Existen variantes del criterio de información anterior:

$$AIC(\beta) = -2 \ln L_X(\beta, S_X(\beta)/n) + 2(p + q + 1)$$

$$\begin{aligned} BIC &= (n - p - q) \ln \left(\frac{n\hat{\sigma}^2}{n - p - q} \right) + n(1 + \ln \sqrt{2\pi}) \\ &\quad + (p + q) \ln \left[\frac{\sum_{t=1}^n X_t^2 - n\hat{\sigma}^2}{p + q} \right], \end{aligned}$$

este último criterio estima consistentemente los ordenes p y q .

Ejercicio 5.8. Veamos primero un ejemplo simulado. Considere que simulamos un proceso ARMA(2,1) en los archivos ARMAEstimation 2.Rmd y ARMAEstimation 2.ipynb. También hay un análisis para datos reales del tipo de interés interbancarios relativo en España y precipitaciones mensuales en Londres .

5.4.2 Comprobación de los supuestos(Diagnóstico basados en los residuales)

Modelo → Apropriado

$$Y_t \underset{Obs}{\longleftrightarrow} \underset{Predicciones}{\hat{Y}_t}$$

Los residuales deben ser consistentes con el modelo.

Sean $\hat{\phi}, \hat{\theta}, \hat{\sigma}^2$ las estimaciones de máxima verosimilitud para un modelo ARMA(p,q), los residuales(*errores de predicción un paso adelante intramuestra*) del modelo se definen como

$$\hat{W}_t = \frac{(X_t - \hat{X}_t(\hat{\phi}, \hat{\theta}))}{(r_{t-1}(\hat{\phi}, \hat{\theta}))^{1/2}} \quad t = 1, 2, \dots, n$$

Si el modelo ARMA(p,q) con los estimadores de máxima verosimilitud es el verdadero modelo que genera a $\{X_t\}$, entonces $\{\hat{W}_t\} \sim RB(0, \hat{\sigma}^2)$.

Sin embargo, solo es posible asumir que X_1, \dots, X_n es generado por un modelo ARMA(p,q) con parámetros desconocidos ϕ, θ, σ^2 cuyas estimaciones son $\hat{\phi}, \hat{\theta}, \hat{\sigma}^2$, con lo cual $\{\hat{W}_t\}$ no es un ruido blanco.

Sin embargo $\{\hat{W}_t, t = 1, \dots, n\}$ deben tener propiedades similares a la sucesión de ruido blanco.

$$W_t(\phi, \theta) = \frac{(X_t - \hat{X}_t(\phi, \theta))}{(r_{t-1}(\phi, \theta))^{1/2}} \quad t = 1, 2, \dots, n$$

Nota 5.9. Note que $E[(W_t(\phi, \theta) - Z_t)^2]$ es pequeño para t grande, con lo cual las propiedades de los residuales $\{\hat{W}_t\}$ deben reflejar las propiedades del ruido blanco $\{Z_t\}$ que genera el proceso ARMA(p, q).

Propiedades que aproximadamente refleja $\{\hat{W}_t\}$

- (i) No correlacionado si $\{Z_t\} \sim RB(0, \hat{\sigma}^2)$
- (ii) Independiente si $\{Z_t\} \sim IID(0, \hat{\sigma}^2)$
- (iii) Distribución normal si $\{Z_t\} \sim N(0, \hat{\sigma}^2)$

Cómo debe ser la gráfica de $\{\hat{W}_t, t = 1, 2, \dots, n\}$?

Debe reflejar el comportamiento de un ruido blanco. Desviaciones de la media cero, por ejemplo por una tendencia, ciclo o fluctuaciones grandes dado que la varianza es no constante, pueden hacer ver que el modelo ajustado no es apropiado.

Autocorrelación muestral de \hat{W}_t

$$\hat{\rho}_w(h) = \frac{\sum_{t=1}^{n-h} [(\hat{W}_t - \bar{W})(\hat{W}_{t+h} - \bar{W})]}{\sum_{t=1}^n (\hat{W}_t - \bar{W})^2} \quad h = 1, 2, \dots$$

$$\bar{W} = \frac{1}{n} \sum_{t=1}^n \hat{W}_t.$$

Sea $\{X_t\}$ un proceso ARMA(p, q) causable e invertible y

$$\hat{\rho}_w = (\hat{\rho}_w(1), \dots, \hat{\rho}_w(h))'$$

para h fijo

$$\hat{\rho}_w \text{ es } AN(0, n^{-1}(T_n - Q))$$

$$Q = T_h \hat{\Gamma}_{p+q}^{-1} T'_h = [q_{ij}]_{i,j=1}^h \quad (5.11)$$

$$T_h = [a_{i-j}]_{\{1 \leq i \leq h, 1 \leq j \leq p+q\}} \quad (5.12)$$

$$\hat{\Gamma}_{p+q} = \left[\sum_{k=0}^{\infty} a_k a_{k+|i-j|} \right]_{i,j=1}^{p+q} \quad (5.13)$$

$$a(z) = (\tilde{\phi}(z))^{-1} = \sum_{j=0}^{+\infty} a_j z^j \quad (5.14)$$

$$\tilde{\phi}(z) = \phi(z)\theta(z) = 1 - \tilde{\phi}_1 z - \dots - \tilde{\phi}_{p+q} z^{p+q}$$

Todo se hace puesto que $\hat{W}_1, \dots, \hat{W}_n$ no es iid. La idea es verificar si las autocorrelaciones $\hat{\rho}_w(h)$ son consecuentes con las propiedades asintóticas mostradas anteriormente. La idea es dibujar las bandas $\pm z_{1-\alpha/2}(1-q_{ii})^{1/2}n^{-1/2}$ un valor por fuera de esas bandas sugieren inconsistencias de los residuales \hat{w}_t con el modelo ajustado.

Pruebas Portmanteau

La idea es utilizar una prueba basada en un estadístico que dependa de los $\hat{\rho}_W(i)$ $1 \leq i \leq h$.

Si el modelo es apropiado la distribución

$$\hat{\rho}_w = (\hat{\rho}_w(1), \dots, \hat{\rho}_w(h))'$$

es aproximadamente

$$N(0, n^{-1}(I_k - T_h(T'_n T_n) T'_h))$$

Así

$$Q = n\hat{\rho}'_w \hat{\rho}_w = n \sum_{j=1}^h \hat{\rho}_w^2(j) \sim \chi^2_{h-(p+q)} \text{ (Box-Pierce)}$$

Luego el modelo no es adecuado si se rechaza al nivel α si

$$Q_w > \chi^2_{1-\alpha; h-p-q}$$

$$\tilde{Q}_w = n(n+2) \sum_{j=1}^h \frac{\hat{\rho}_w^2(j)}{n-j} \sim \chi^2_{h-(p+q)} \text{ Mejor aproximación Ljung-Box}$$

$$\tilde{Q}_{ww} = n(n+2) \sum_{j=1}^h \frac{\hat{\rho}_{ww}^2(j)}{n-j} \sim \chi^2_h \text{ MCleod=LI}$$

* Pruebas de aleatoriedad

* Pruebas de Normalidad

* Shapiro-Wilk

$$W = \frac{\left(\sum_{i=1}^T a_i X_{(i)}\right)^2}{\sum_{i=1}^n (X_i - \bar{X})^2} \quad (a_1, \dots, a_n) = \frac{m' v^{-1}}{(m' v^{-1} v^{-1} m)^{-1/2}}$$

$m = (m_1, \dots, m_T)$ v Esperanza de los estadísticos de orden respecto a la varianza

* Jarque Bera

$$JB = \frac{n-T-1}{6} \left(S^2 + \frac{1}{4}(C-3)^2 \right) \quad JB > \chi_2^2$$

S : Asimetría C : Kurtosis

* Pruebas de Estabilidad basadas en las sumas de cuadrados(cusum's)
Si \hat{W}_t son los residuales, entonces la estadística cusum se computa

$$W_t = \sum_{r=k+1}^t \hat{W}_r / s \text{ para } t = k+1, \dots, n$$

donde s es la desviación estándar de los residuales \hat{W}_t . Se hace la gráfica de W_t contra t .

Las bandas de confianza del 95% son líneas que se unen en los siguientes puntos $[k, \pm 0.948\sqrt{n-k}]$ y $[n, \pm 3 \times 0.948\sqrt{n-k}]$

La estadística cusumsq se computa como sigue

$$S_t = \frac{\sum_{r=k+1}^t \hat{W}_r^2}{\sum_{r=k+1}^n \hat{W}_r^2}.$$

Lo cual permite hacer una gráfica de S_t contra t , junto con un par de líneas paralelas $s_r = \pm c_0 + \frac{r-k}{T-k}$, donde c_0 depende del nivel de significancia y del tamaño de la muestra. Los valores de los cuantiles para elaborar las gráficas cusum's se pueden encontrar en los artículos [4] y [10].

6

Modelo ARIMA para Series de Tiempo No Estacionarias

Muchas series de tiempo son realizaciones de proceso estocásticos no estacionarias. Hay algunas razones por las cuales una serie de tiempo es no estacionaria, una de ellas es la presencia de una tendencia(determinística o estocástica), esto hace que la media del proceso no sea constante por ejemplo. Otra razón de no estacionariedad en una serie de tiempo es debido a la presencia de una componente cíclica(estacional o no estacional), esto también puede implicar que la media del proceso no sea constante. Finalmente, si la serie parece tener un rango de valores diferente a lo largo del tiempo, entonces parece ser que la varianza marginal tampoco es constante a lo largo del tiempo, y esto implica la serie es no estacionaria. La serie de tiempo puede presentar una o varias de estas características, lo cual la hace no estacionaria. En éste capítulo hablaremos de un tipo de modelo para analizar series de tiempo con tendencia. La tendencia que se considerará será estocástica y se presenta cuando hay presencia de una raíz unitaria en el polinomio autoregresivo. El modelo ARIMA es introducido incluir la presencia de una raíz unitaria en la serie. Debido a que la tendencia estocástica no es la única forma de tendencia que puede presentarse en una serie de tiempo, hay mas formas de hacer el modelamiento de una serie de tiempo con tendencia. En el siguiente capítulo veremos metodologías alternativas para el modelamiento de series de tiempo con tendencia y los modelos ARIMA estacionales(SARIMA).

6.1 Modelo Autoregresivos Integrados y de Promedio Móviles(ARIMA)

Un modelo ARIMA no es mas que la extensión del modelo ARMA donde se admite al menos una raíz en $z = 1$ en el polinomio autoregresivo $\phi(z)$. Del capítulo 3, tenemos que el teorema 3.19 nos menciona las condiciones para que la ecuación ARMA tenga solución estacionaria. Si el polinomio $\phi(z)$ tiene una raíz en $z = 1$, eso implicará que la ecuación no tendrá una solución estacionaria, por lo tanto, sus realizaciones no tendrán trayectorias estables, es decir, producirán series de tiempo no estacionarias. En seguida daremos la definición de los modelo ARIMA.

Definición 6.1. *Si d es un entero no negativo, entonces $\{X_t\}$ es llamado un proceso ARIMA(p,d,q) si el proceso $Y_t = (1 - B)^d X_t$ es un proceso ARMA causal. Esta definición es equivalente a decir que $\{X_t\}$ satisface la ecuación en diferencias*

$$\phi^*(B)X_t = \phi(B)(1 - B)^d X_t = \theta(B)Z_t \quad Z_t \sim RB(0, \sigma^2) \quad (6.1)$$

$\phi(z)$ y $\theta(z)$ son polinomios de grados p y q respectivamente y $\phi(z) \neq 0$ tal que $|z| \leq 1$.

Veamos una ejemplo simulado de un proceso ARIMA en python usando el archivo Arima2.ipynb.

En la gráfica 6.1 podemos ver una serie simulada de un proceso ARIMA. Note que la serie es no estacionaria y presenta una tendencia que hemos llamado estocástica. Adicionalmente podemos como su función de autocorrelación simple decae lentamente en la figura 6.2, la cual es otra característica de un proceso ARIMA.

Note también que al diferenciar la serie simulada una vez, es decir $(X_t - X_{t-1})$, la serie resultante ya no presenta tendencia, vea la figura 6.3.

Una limitación de los procesos ARIMA, es que el tipo de no estacionariedad que se está modelando es aquella donde tenemos la presencia de una raíz unitaria de multiplicidad d . Sin embargo esta no es la única forma de no estacionariedad. Supongamos que en el modelo

$$\phi^*(B)X_t = \theta(B)Z_t$$

el polinomio $\phi^*(z)$ tiene un cero en $z = e^{i\theta}$ con $\theta \in (-\pi, \pi]$ e i es el complejo $i = \sqrt{-1}$. No es difícil verificar que la norma de esta raíz es 1. Para esto se

6.1. MODELO AUTOREGRESIVOS INTEGRADOS Y DE PROMEDIO MÓVILES(ARIM)

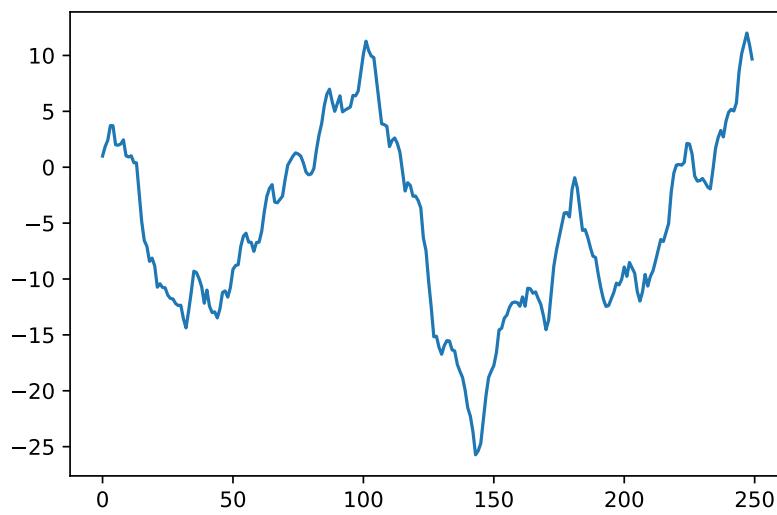


Figura 6.1: Gráfico de una Serie de tiempo ARIMA usando Python

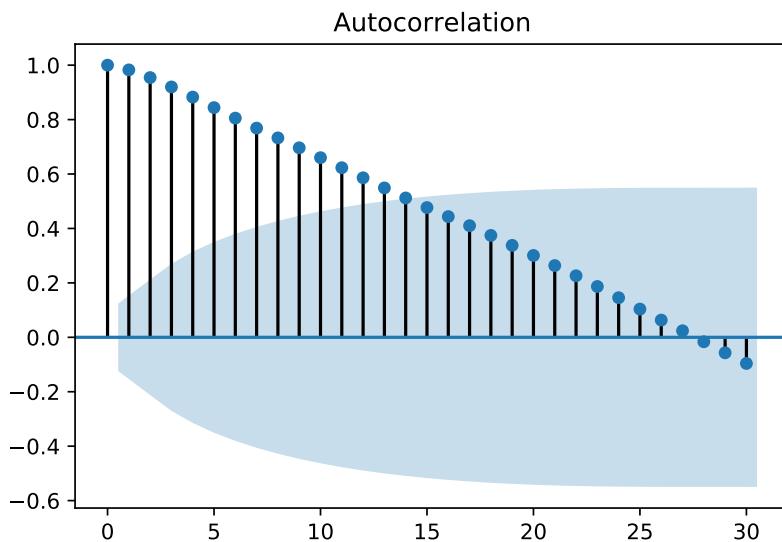


Figura 6.2: Gráfica de la ACF de una serie de tiempo simulada de un proceso ARIMA(1,1,0) con $\phi = 0.5$

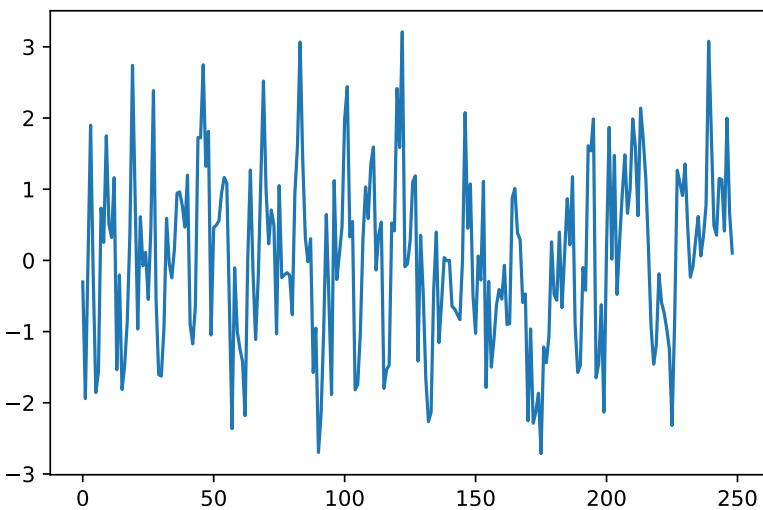


Figura 6.3: Gráfica de serie diferenciada

usa el hecho que cualquier complejo se puede escribir (usando la fórmula de Euler) de la forma

$$e^{ir\theta} = r(\cos \theta + i \sin \theta)$$

y así $|e^{ir\theta}|^2 = (r(\cos \theta + i \sin \theta))(r(\cos \theta - i \sin \theta)) = r^2(\cos^2 \theta + \sin^2 \theta) = r^2$. Por otro lado, para que nuestro proceso sea de valor real, es necesario incluir la raíz compleja y su conjugado complejo, es decir consideremos que el polinomio $\phi^*(B) = (1 - e^{i\theta}B)(1 - e^{-i\theta}B)$, el cual puede verificarse que es equivalente a $\phi^*(B) = 1 - 2 \cos \theta B + B^2$ usando la fórmula Euler para escribir un número complejo. Usaremos esto para simular un proceso bajo estas condiciones asumiendo que $\theta = \frac{\pi}{3}$. En las gráficas 6.4 y 6.5 podemos observar la serie simulada y su respectivo ACF muestral, para un proceso con raíces unitarias complejas. Note que en este caso la serie no presenta tendencia pero si un comportamiento cíclico, y además su función de autocorrelación simple presenta un comportamiento oscilatorio con decaimiento lento.

Note que en la figura 6.6 se encuentra la simulación de una caminata aleatoria y de una caminata aleatoria con drfit.

En la gráfica 6.7 podemos ver dos series, una simulada por un proceso ARIMA con drfit y la otra mediante un proceso que presenta tendencia determinística a la cual se le suma un ruido. La serie simulada por un proceso

6.1. MODELO AUTOREGRESIVOS INTEGRADOS Y DE PROMEDIO MÓVILES(ARIM)

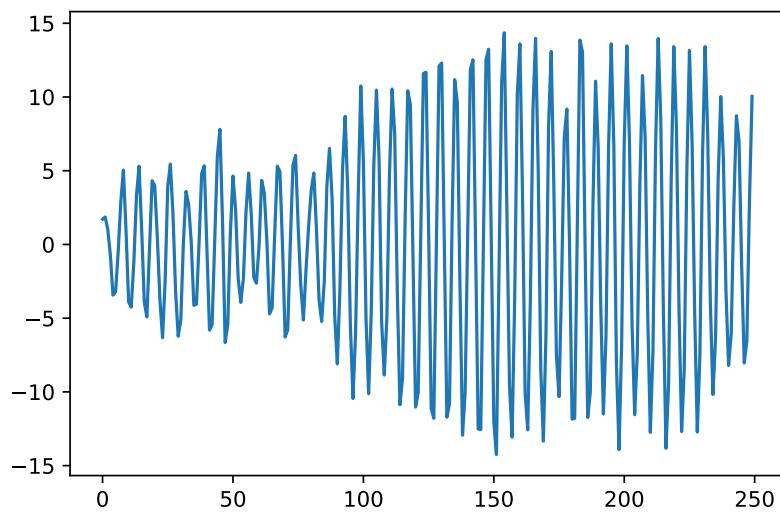


Figura 6.4: Simulación de un proceso con raíces unitarias complejas .

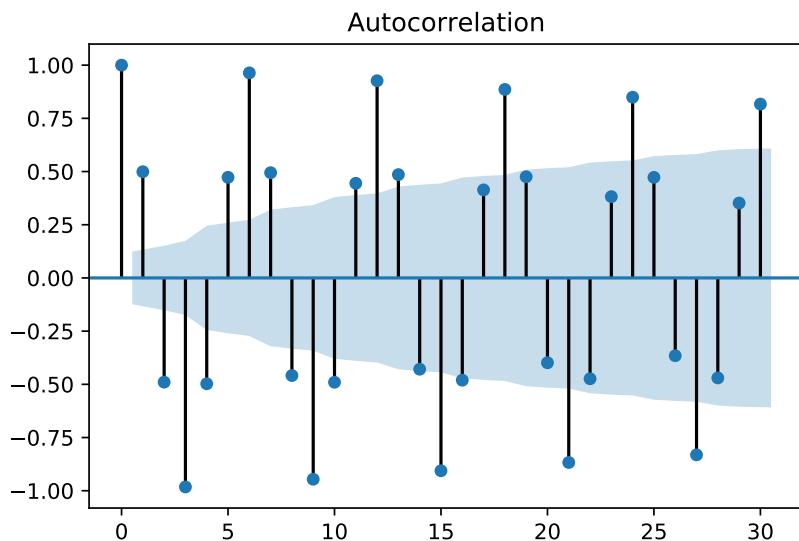


Figura 6.5: ACF de un proceso simulado con raíces unitarias complejas sobre el círculo unitario.

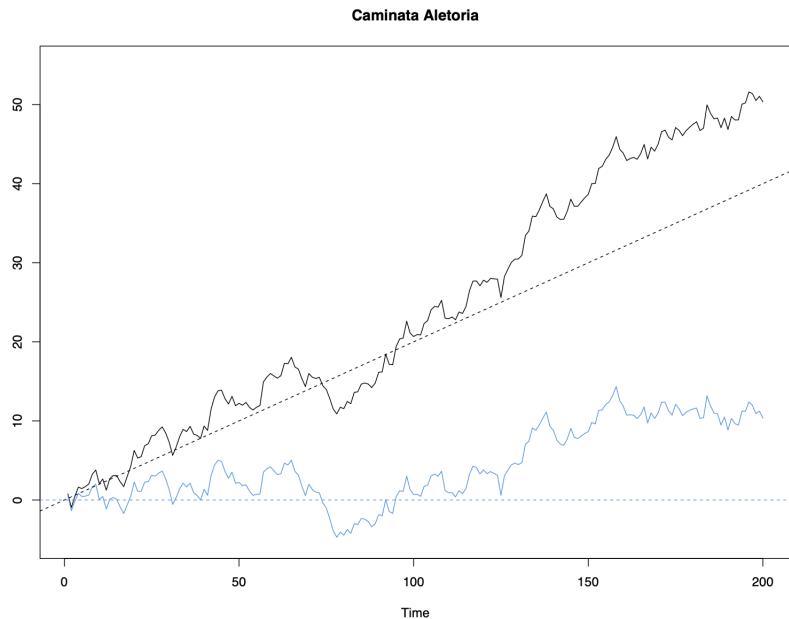


Figura 6.6: Caminata Aleatoria con y sin Drift

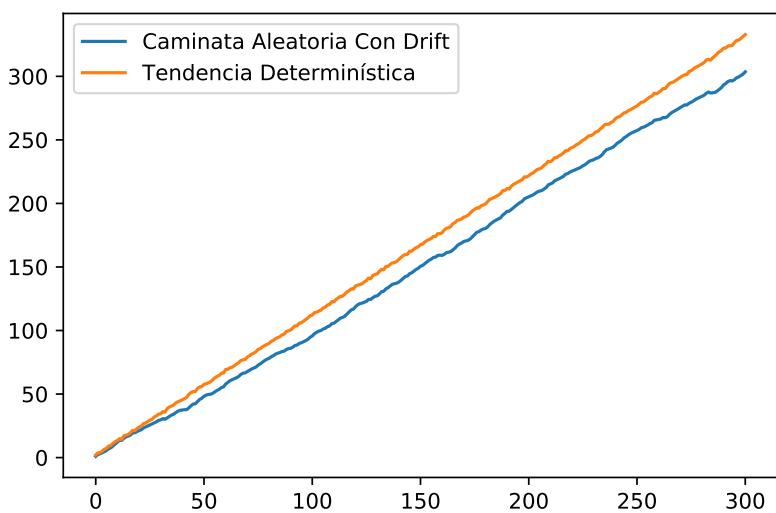


Figura 6.7:

6.1. MODELO AUTOREGRESIVOS INTEGRADOS Y DE PROMEDIO MÓVILES(ARIM)

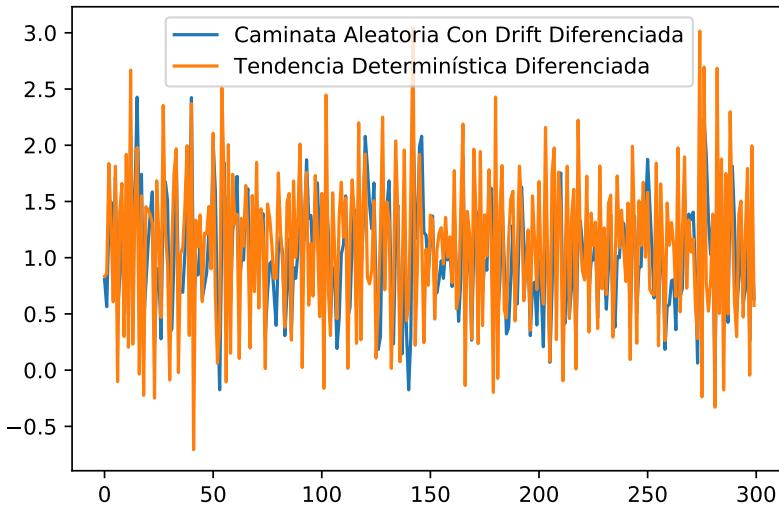


Figura 6.8: Series con tendencia estocástica y determinísticas diferenciadas

ARIMA con drfit es obtenida añadiendo a la ecuación 6.1 una constante c , es decir

$$\phi(B)(1 - B)^d X_t = c + \theta(B)Z_t.$$

La ecuación de un modelo con tendencia determinística es de la forma

$$X_t = a + bt + Z_t.$$

Note no se puede ver la diferencia de una serie con tendencia estocástica con drift y otra con tendencia determinística viendo solo la gráfica de la serie. En la gráfica 6.8 podemos ver que una serie con tendencia determinística o con tendencia estocástica pueden transformarse a estacionarias con diferenciarlas.

Ejemplo 6.2. ARIMA(1,1,0)

Consideremos el proceso ARIMA(1,1,0) tal que

$$(1 - \phi B)(1 - B)X_t = Z_t \quad Z_t \sim RB(0, \sigma^2).$$

Entonces se puede verificar que

$$X_t = X_0 + \sum_{j=1}^t Y_j \quad t \geq 1.$$

En efecto, sea $Y_t = (1 - B)X_t = \sum_{j=0}^{\infty} \phi^j Z_{t-j}$ y note que

$$X_t = X_0 + (X_1 - X_0) + (X_2 - X_1) + (X_3 - X_2) + \cdots + (X_t - X_{t-1})$$

con lo cual podemos escribir

$$X_t = X_0 + (1 - B)X_1 + (1 - B)X_2 + (1 - B)X_3 + \cdots + (1 - B)X_t$$

y así

$$X_t = X_0 + Y_1 + Y_2 + Y_3 + \cdots + Y_t$$

y finalmente

$$X_t = X_0 + \sum_{j=1}^t Y_t$$

6.2 Modelamiento de Series no Estacionarias Vía Modelos ARIMA

Lo primero que se debe hacer antes de ajustar un modelo ARIMA es verificar la presencia de una raíz unitaria en el polinomio autoregresivo. Explicaremos en seguida un prueba de raíz unitaria propuesta por Dickey y Fuller en 1979. Particularmente, sean x_1, \dots, x_n observaciones de un modelo AR(1),

$$X_t - \mu = \phi_1(X_{t-1} - \mu) + Z_t \quad Z_t \sim RB(0, \sigma^2), \quad (6.2)$$

donde $|\phi_1| < 1$ y $\mu = E[X_t]$. Para n suficientemente grande

$$\hat{\phi}_1 \sim N(\phi_1, (1 - \phi_1)/n).$$

Si $\phi_1 = 1$, esta aproximación no es aplicable y no puede ser usada para probar a

$$H_0 : \phi_1 = 1 \quad vs \quad H_1 : \phi_1 \neq 1.$$

Para crear una prueba, la idea es escribir el modelo 6.2 como:

$$\nabla X_t = X_t - X_{t-1} = \phi_0^* + \phi_1^* X_{t-1} + Z_t \quad \{Z_t\} \sim RB(0, \sigma^2),$$

donde $\phi_0^* = \mu(1 - \phi_1)$ y $\phi_1^* = 1 - \phi_1$. Ahora, sea $\hat{\phi}_1^*$ el estimador de mínimos cuadrados de ϕ_1^* de la regresión entre ∇X_t sobre 1 y X_{t-1} .

Luego el error estándar de $\hat{\phi}_1^*$ es

$$\hat{SE}(\hat{\phi}_1^*) = S \left(\sum_{t=2}^n (X_{t-1} - \bar{X})^2 \right)^{-1/2}$$

$$S = \frac{\sum_{t=2}^n (\nabla X_t - \hat{\phi}_0^* - \hat{\phi}_1^* X_{t-1})^2}{n - 3}$$

Dickey y Fuller mostraron que

$$\hat{\tau}_\mu = \frac{\hat{\phi}^*}{\hat{SE}(\hat{\phi}_1^*)}$$

Tiene una distribución límite no estándar bajo $H_0 : \phi_1^* = 0$, es decir, $H_0 : \phi_1 = 1$. Esta es conocida como la prueba de Dickey-Fuller aumentada (ADF). Lo de aumentada es porque se permite que en la prueba original sean incluidas diferencias rezagadas en la ecuación de regresión para modelos más generales.

Se rechaza H_0 si $\hat{\tau}_M < \tau \rightarrow$ cuantil de la distribución límite.

Los cuantiles son los siguientes:

α	0.01	0.05	0.10
τ	-3.43	-2.86	-2.57

Note que la prueba de Dickey-Fuller es una prueba a una cola.

Ejercicio 6.3. Simule una serie ARIMA(1,1,0) con Drfit de tamaño $T = 300$ y $\phi = 0.5$. Ajuste un modelo ARIMA con drift y un modelo AR(1) con una tendencia determinística(Es decir debe crear dos covariables o variables regresoras y añadirlas al modelo). Compare los criterios de información para los dos modelos. Haga Rolling para ver que modelo hace mejores pronósticos, usando como conjunto de entrenamiento $T = 200$ para las predicciones 1 y 2 pasos adelante.

Repita el ejercicio anterior, pero ahora simule los datos de un modelo AR(1) con tendencia determinística.

Repita la experiencia anterior 500 veces y almacene los errores cuadráticos medios encontrados con el Rolling. Qué conclusión puede dar de los resultados de simulación?

6.3 Pronósticos con Modelos ARIMA

Veremos como generalizar para obtener los pronósticos en Modelos ARIMA. Si $d \geq 1$, los momentos $E[X_t]$ y $E[X_t X_{t+h}]$ no pueden ser determinadas por la ecuación en diferencias

$$\phi^*(B)X_t = \phi(B)(1 - B)^d X_t = \theta(B)Z_t \quad \{Z_t\} \sim RB(0, \sigma^2),$$

con lo cual los predictores lineales para $\{X_t\}$ no pueden ser obtenidos como en el caso de los modelos ARMA sin otras condiciones.

Sea $\{Y_t\}$ un proceso ARMA(p,q) causable y X_0 cualquier variable aleatoria Y definamos

$$X_t = X_0 + \sum_{j=1}^t Y_j \quad t = 1, 2, \dots$$

Así $\{X_t\}$ es un ARIMA(p,1,q) con media

$$E[X_t] = E[X_0]$$

Y covarianzas

$$E[X_t X_{t+h}] - (E[X_0])^2$$

Las cuales dependen de $Var(X_0)$ y $Cov(X_0, Y_j)$.

El mejor predictor de X_{n+1} basados en X_0, X_1, \dots, X_n , es $P_{S_n}X_{n+1}$

$$S_n = \overline{sp}\{X_0, X_1, \dots, X_n\} = \overline{sp}\{X_0, Y_1, \dots, Y_n\}$$

Así

$$P_{S_n}X_{n+1} = P_{S_n}(X_0 + Y_1 + \dots + Y_n) = X_n + P_{S_n}Y_{n+1}$$

$P_{S_n}Y_{n+1}$ depende de $E[X_0Y_j]$; $E[X_0^2]$, pero si X_0 es no correlacionada con Y_j entonces $P_{S_n}Y_{n+1}$ es obtenido proyectando sobre $\overline{sp}\{Y_1, \dots, Y_n\}$ únicamente y se puede usar, por lo tanto los métodos para hacer pronóstico en modelos ARMA(p,q).

Para el caso más general, es decir asumamos que

$$(1 - B)^d X_t = Y_t \quad t = 1, 2, \dots$$

$\{Y_t\}$ es un proceso ARMA(p,q) causable y el vector (X_{1-d}, \dots, X_0) es no correlacionado con Y_t , $t > 0$, luego

$$X_t = Y_t - \sum_{j=1}^d \binom{d}{j} (-1)^j X_{t-j} \quad t = 1, 2, \dots$$

La idea es computar

$$P_{S_n}X_{n+h} = P_{\overline{sp}\{X_{1-d}, X_{2-d}, \dots, X_n\}}X_{n+h}$$

Sea

$$P_n Y_{n+h} = P_{\overline{sp}\{Y_1, \dots, Y_n\}} Y_{n+h}$$

con lo cual $\hat{Y}_{n+1} = P_n Y_{n+1}$. Por lo tanto

$$S_n = \overline{sp}\{X_1 - d_1, \dots, X_0, Y_1, Y_n\}$$

y bajo el supuesto

$$\overline{sp}\{X_{1-d}, \dots, X_0\}_{S_0} \perp \overline{sp}\{Y_1, \dots, Y_n\}$$

entonces

$$P_n Y_{n+h} - P_{S_0} Y_{n+h} + P_n Y_{n+h} = P_n Y_{n+h}$$

luego con $t = n + h$ y dado que

$$X_t = Y_t - \sum_{j=1}^d \binom{d}{j} (-1)^j X_{t-j} \quad t = 1, 2, \dots$$

$$P_{S_n} X_{n+h} = P_n Y_{n+h} - \sum_{j=1}^d \binom{d}{j} (-1)^j P_{S_n} X_{n+h-j}$$

Luego $P_{S_n} X_{n+1}, P_{S_n} X_{n+2}, \dots$ se pueden computar en forma recurrente también de la ecuación anterior.

En términos del algoritmo de innovaciones tenemos que

$$\sigma^2(h) = E[(X_{n+h} - P_{S_n} X_{n+h})^2] = \sum_{j=0}^{h-1} \left(\sum_{r=0}^j \chi_r \theta_{n+h-r-1,j-r} \right)^2 \nu_{n+h-j-1}$$

$$\chi(z) = \sum_{r=0}^{\infty} \chi_r z_r = (1 - \phi_1^* z - \cdots - \phi_{p+d} z^{p+d})^{-1} \quad |z| \leq 1$$

$$v_{n+h-j-1} = E[(X_{n+h-j} - X_{n+h-j}^*)^2] = E[(Y_{n+h-j} - \hat{Y}_{n+h-j})^2]$$

Para n suficientemente grande una aproximación para $\sigma_n^2(h)$ esta de

$$\sum_{j=0}^{h-1} \psi_j^2 \sigma^2$$

$$\psi(z) = \sum_{j=0}^{\infty} \psi_j z^j = (\phi^*(z))^{-1} \theta(z) \quad |z| \leq 1.$$

Vale la pena decir que la estimación de los modelos ARIMA se llevan a cabo vía máxima verosimilitud de manera análoga como en los modelos ARMA. Se mantienen las distribuciones límite y las propiedades de los estimadores.

6.4 Series de Tiempo Estacionales y Modelos SARIMA

En ocasiones, las series de tiempo presentan patrones cíclicos de periodo fijo. Por ejemplo, puede que las ventas de un almacén tengan un crecimiento todos los meses de junio y diciembre, y un posible decrecimiento en los meses de enero y febrero a lo largo de todos los años. Es decir, tenemos picos y valles en los mismo meses, a lo largo de todos los años. Note por ejemplo que en la serie de pasajeros se presenta éste fenómeno. También en la serie 1.1, podemos ver picos bajos de lluvias en los meses diciembre y enero, al igual que en los meses de julio y agosto, mientras que vemos un pico alto de lluvias en los meses de marzo y abril, al igual que en los meses de octubre y noviembre a lo largo de todos los años. Este fenómeno se conoce

como *estacionalidad*, y se representa formalmente como que el proceso $\{X_t\}$ satisface

$$E[X_t] = E[X_{t+s}]$$

para cada tiempo t y un entero s positivo llamado periodo. Éste capítulo se desarrollará siguiendo las pautas del libro [29]. Note que una serie que presente estacionalidad es no estacionaria. El periodo estacional, s , se define como número de observaciones que forman el ciclo estacional. Por ahora, suponemos que hay sólo existe un sólo tipo de estacionalidad. Porque por ejemplo en datos diarios podría existir una estacionalidad semana, con $s = 7$, otra mensual con $s = 30$, y otra anual, con $s=365$.

Hay varias formas de introducir la estacionalidad en una serie de tiempo (estocástica y no estocástica). La forma mas simple que puede es a través de un efecto constante que se suma a los valores de la serie. Supongamos por ejemplo que la serie es una suma de una componente estacional $S_t^{(s)}$, y un proceso estacionario, n_t , así que el modelo para la serie observable Z_t será:

$$Z_t = S_t^{(s)} + n_t. \quad (6.3)$$

El proceso 6.3 es no estacionario ya que

$$E[Z_t] = E[S_t^{(s)}] + \mu,$$

donde μ es la media del proceso n_t . Note que como la componente estacional no toma el mismo valor en todos los periodos por definición 2.7. Ver ejemplo Serie de número de pasajeros, Ozono y lluvia.

Se puede considerar distintas hipótesis respecto al comportamiento del proceso estacional $S_t^{(s)}$. Por ejemplo que $S_t^{(s)}$ sea un poceso determinístico o determinista, es decir, una función constante para el mismo mes en distintos años:

$$S_t^{(s)} = S_{t+ks}^{(s)}, k = \pm 1, \pm 2, \dots,$$

como por ejemplo, los coeficientes estacionales pueden seguir una función sinusoidal. Estas funciones serán poco eficaces cuando la estacionalidad siga un patrón determinista pero no sinusoidal, por ejemplo en series de producción, se muestran valores constantes todos los meses excepto en los meses de vacaciones, digamos diciembre y enero, así que le valor esperado es mas bajo en esos meses. La idea en estos casos es introducir $11(s - 1)$ variables

impulso o dummy $I_t^{(j)}$, que tomen 1 en un mes cero en el resto, es decir:

$$Z_t = \mu + \sum_{j=1}^{s-1} w_j I_t^{(j)} + n_t.$$

Otras serie pueden no presentar determinista, sino que la componente estacional evoluciona o tiene dinámica propia con respecto al tiempo. Lo cual supone que sigue un proceso estocástico, es decir, los factores estacionales no son constantes, pero siguen un proceso estacionario, oscilando alrededor de un valor medio de acuerdo a:

$$S_t^{(s)} = \mu^{(s)} + \eta_t,$$

donde $\mu^{(s)}$ es una constante que depende del mes y representa el efecto determinista de la estacionalidad y η_t es un proceso estocástico estacionario de media cero que introduce la variabilidad de cada año.

La otra forma de representar la estacionalidad consiste en que esta sea cambiante en el tiempo sin ningún valor medio fijo, es decir la estacionalidad sigue un proceso estocástico no estacionario, el cual se puede asumir que evoluciona como una caminata aleatoria:

$$S_t^{(s)} = S_{t-s}^{(s)} + \eta_t.$$

Por supuesto, pueden haber mas tipos de estacionalidad, las cuales pueden seguir cualquier proceso estocástico no estacionario. Sin embargo, vale la pena decir, que en cualquiera de los tres casos presentados anteriormente, podemos convertir una serie estacional en estacionaria aplicando diferencia estacional ver [29] página 205. Definimos el operador diferencia estacional de periodo s como:

$$\nabla_s = 1 - B^s,$$

es decir, $\nabla_s Z_t = Z_t - Z_{t-s}$. Sin embargo, para éste capítulo se considerará la estacionalidad de tipo estocástica y vendrá como una generalización del modelo ARMA de forma análoga a como se hizo con el modelo ARIMA para incluir la presencia de una raíz unitaria estacional en el polinomio autorregresivo, es decir, la estacionalidad está dada por la inclusión del operador diferencia estacional

$$(1 - B^s)^D.$$

Además de incorporar la dependencia regular, que es asociada a los intervalos de medida de la serie, la dependencia estacional, que es asociada a observaciones separadas por s periodos también será incorporada. Note que

si $D = 1$, entonces tenemos que el operador diferencia estacional aplicado a un proceso $\{X_t\}$ quiere decir que $(1 - B^s)X_t = X_t - X_{t-s}$ y no es lo mismo que aplicar el operador diferencia ordinaria s veces, $(1 - B)^s X_t$. Veamos un proceso simulado con la presencia de una raíz unitaria en R usando el script Modelo SARIMA.R.

Un enfoque para incorporar todas las componentes de dependencia (regular y estacional), de tendencia y estacional es a través de un modelo multiplicativo, es decir, modelar separadamente la dependencia regular y la estacional.

Definición 6.4. Si d y D son enteros no negativos, entonces $\{X_t\}$ se llama un proceso estacional ARIMA(p, d, q) \times (P, D, Q)_s o SARIMA(p, d, q) \times (P, D, Q)_s con periodo s , si la serie diferenciada $Y_t = (1 - B)^d(1 - B^s)^D X_t$ es causable, es decir, la ecuación en diferencias estocástica

$$\phi(B)\Phi(B^s)Y_t = \theta(B)\Theta(B^s)Z_t \quad \{Z_t\} \sim RB(0, \sigma^2),$$

tiene una solución causal. Los polinomios $\phi(B)$ y $\theta(B)$ son llamados los polinomio autoregresivos y de promedios móviles ordinarios o regulares, mientras que los polinomios $\Theta(B^s) = 1 + \Theta_1 B^s + \dots + \Theta_P B^{sQ}$ y $\Phi(B^s) = 1 - \Phi_1 B^s - \dots - \Phi_P B^{sP}$ son llamados los polinomios autoregresivos y de promedio móviles estacionales. D es llamado el orden de integración estacional.

Una forma de entender la construcción de éste modelo es como sigue, para una serie mensual de periodo $s = 12$ a lo largo de r años:

		Mes			
Año		1	2	...	12
1	X_1	X_2	...	X_{12}	
2	X_{13}	X_{14}	...	X_{24}	
3	X_{25}	X_{26}	...	X_{36}	
:	:	:	⋮	⋮	⋮
r	$X_{1+12(r-1)}$	$X_{2+12(r-1)}$...	$X_{12+12(r-1)}$	
	Serie 1	Serie 2	...	Serie 12	

es decir, para el mes j -ésimo, X_{j+12t} , $t = 0, 1, \dots, r-1$, satisface la ecuación en diferencias ARMA(P, Q)

$$X_{j+12t} = \Phi_1 X_{j+12(t-1)} + \dots + \Phi_p X_{j+12(t+p)} + U_{j+12(t-1)} + \dots + \Theta_q U_{j+12(t-Q)} (*)$$

donde $\{U_{j+12r}, t = \dots, 0, -1, 0, 1, \dots\} \sim RB(0, \sigma_U^2)$.

Podemos escribir en forma compacta el modelo (*)

$$\Phi(B^{12})X_t = \Theta(B^{12}U_t) \quad \text{modelo entre años}$$

$$\Phi(z) = 1 - \Phi_1 z - \dots - \Phi_p z^p$$

$$\Theta(z) = 1 - \Theta_1 z - \dots - \Theta_q z^q$$

$\{U_{j+12t}, t = \dots, -1, 0, 1, \dots\} \sim RB(0, \sigma_U^2)$ para cada j.

Para involucrar la dependencia entre las 12 series, se asume que $\{U_t\}$ sigue un modelo ARMA(p,q) esto implica

$$(**)\phi(B)U_t = \theta(B)Z_t,$$

con $\{Z_t\} \sim RB(0, \sigma^2)$. Esto implica

- ◊ No correlación o correlación cero entre dos valores consecutivas U_t .
- ◊ Correlación dentro las 12 sucesiones $\{U_{j+12t}, t = \dots, -1, 0, 1, \dots\}$.

Lo que sucede entonces es que $E[U_t, U_{t+12j}]$ es pequeño para $j = \pm 1, \pm 2, \dots$. Se pueden permitir diferencias ordinarias y estacionales dentro de los modelos (*) y (**) lo cual se obtiene la siguiente definición del modelo SARIMA 6.4 combinando los modelos.

Nota 6.5. La interacción de los dos modelos describe la estructura de dependencia entre año y dentro de año la cual no es fácil describir el comportamiento, lo que hace difícil la identificación del modelo.

¿Cómo entonces identificarlos?

- ✓ Hacer estabilización de la varianza marginal usando transformación Box-Cox si es necesario.
- ✓ Encontrar d y D tal que la serie

$$y_t = (1 - B)^d (1 - B^s)^D x_t$$

tenga la apariencia de estacionario.

- ✓ Examinar la autocorrelación y autocorrelación parcial de $\{Y_t\}$ en los rezagos que son múltiplos de s , con el objeto de identificar las ordenes P y Q , es decir, usar $\hat{\rho}(ks)$ y $\hat{\alpha}(ks)$ con $k = 1, 2, \dots$. Las órdenes p y q son seleccionadas usando $\hat{\rho}(1), \hat{\rho}(2), \dots, \hat{\rho}(s-1)$ y $\hat{\alpha}(1), \hat{\alpha}(2), \dots, \hat{\alpha}(s-1)$ complementar la selección usando los criterios de información.
- ✓ Posteriormente se procede a estimar los parámetros $\phi, \theta, \Phi, \Theta, \sigma^2$ basados en la serie original $\{x_t\}$.
- ✓ Hacer pronósticos de manera similar a los modelos ARIMA.
- ✓ Verificación de supuestos.

Los pronósticos de un modelo ARIMA pueden ser aplicados a procesos ARIMA estacionales como es señalado en [2]. Primero se hace la predicción de un modelo ARMA $\{Y_t\}$ donde $Y_t = (1 - B)^d(1 - B^s)^D X_t$ como en la sección 3.5.1. Después se expande el operador $(1 - B)^d(1 - B^s)^D$ para obtener una ecuación análoga a la obtenida en la sección 6.3 en modelos ARIMA. Adiconalmente, vale la pena decir que la estimación de los modelos SARIMA se llevan a cabo vía máxima verosimilitud de manera análoga como en los modelos ARIMA. Se mantienen las distribuciones límite y las propiedades de los estimadores.

Metodologías Alternativas para Series de Tiempo No Estacionarias

En éste capítulo vamos a considerar alternativas al modelo SARIMA para el modelamiento de series de tiempo no estacionarias. Entre las alternativas que se abordarán están: descomposición basados en filtros, suavizamiento exponencial, modelos estructurales de series de tiempo y metodologías basadas en redes neuronales y árboles de decisión. Las primeras tres metodologías se basarán en el supuesto que una serie de tiempo observable puede descompuesta en una componente de tendencia y una componente estacional, es decir, se $\{X_t\}$ puede descomponerse de la siguiente forma aditiva

$$X_t = m_t + S_t + Y_t, \quad (7.1)$$

donde m_t : función que cambia suavemente,

S_t : función de periodo conocido d,

Y_t : ruido aleatorio estacionario en el sentido débil.

Un modelo multiplicativo puede ser considerado como modelo alternativo al aditivo,

$$X_t = m_t \times S_t \times Y_t, \quad (7.2)$$

el cual es equivalente al modelo aditivo si aplicamos una trasformación logarítmica. Una componente adicional puede ser incluida en el modelo la cual será llamada componente cíclica, la cual es diferente a la componente estacional. Si bien, los modelos basados en redes neuronales no son del todo modelos estadísticos, estos serán introducidos dada su importancia en lo que hoy se llama Data Science y su utilidad cuando se tiene una gran cantidad de datos a la mano.

7.1 Descomposición por Filtros de Promedios Móviles

Vamos a empezar con el modelo más simple, el cual es un modelo que sólo tiene la componente de tendencia. Consideremos el modelo $X_t = m_t + Y_t$, donde $E[Y_t] = 0$.

Suavizado con un filtro de promedio móvil. Sea q un entero no negativo y consideremos el promedio móvil a dos colas

$$W_t = (2q+1)^{-1} \sum_{j=-q}^q X_{t-j}$$

del proceso $\{X_t\}$.

Se puede verificar que para $q+1 \leq t \leq n-q$ $W_t \approx m_t$ basados en que m_t es aproximadamente lineal para $[t-q, t+q]$. Por lo tanto,

$$\hat{m}_t = (2q+1)^{-1} \sum_{j=-q}^q X_{t-j}, \quad q+1 \leq t \leq n-q.$$

Nota 7.1. Aquí los coeficientes del filtro $\hat{m}_t = \sum_{j=-\infty}^{\infty} a_j X_{t-j}$ son $a_j = (2q+1)^{-1}$ para $j = -q, \dots, q$ y cero en otro caso.

Considere ahora que

$$X_t = m_t + S_t + Y_t, \quad E[Y_t] = 0, S_{t+d} = S_t \text{ y } \sum_{j=1}^d S_j = 0.$$

Procedemos ahora de la siguiente manera:

Paso 1) Estimar la tendencia aplicando un filtro de promedios móviles con el objeto de eliminar la tendencia. Si $d = 2q$ entonces

$$\hat{m}_t = (0.5x_{t-q} + x_{t-q+1} + \dots + x_{t+q-1} + 0.5x_{t+q})/d \quad q < t \leq n-q.$$

Si $d = 2q+1$, se usa el mismo filtro con una modificación.

Paso 2) Estimar la componente estacional. Para cada $k = 1, \dots, d$, computemos el promedio w_k de las desviaciones $\{(x_{k+jd} - \hat{m}_{k+jd}), q < k+jd \leq n-q\}$.

Así, $\hat{S}_k = w_k - d^{-1} \sum_{j=1}^d w_j$ $k = 1, \dots, d$, y $\hat{S}_k = \hat{S}_{k-d}$, $k > d$. Ahora, los datos son "desestacionalizados"

$$d_t = x_t - \hat{S}_t, \quad t = 1, 2, \dots, n,$$

y la componente de tendencia m_t es estimada de nuevo basados en d_t como antes. al final $\hat{Y}_t = x_t - \hat{m}_t - \hat{S}_t \quad t = 1, 2, \dots, n$.

Note que hay mas tipos de descomposiciones, por ejemplo está la descomposición *X11* desarrollado por la Oficina del Censo de EE. UU. y Estadísticas de Canadá; también está la descomposición *SEATS* desarrollada por el banco de España. Finalmente tenemos la descomposición *STL*, de las siglas en inglés *Seasonal and Trend decomposition using Loess*, el cual usa la técnica *Loess* que permite estimar relaciones no lineales de forma no paramétrica. Supongamos que se tienen observaciones x_i y y_i para $i = 1, \dots, n$; la curva de regresión *Loess* $\hat{g}(x)$, es una suavizamiento de y dado x , que se computa como sigue:

- 1.) Elija un entero positivo $q \leq n$.
- 2.) Se seleccionan los q valores de x_i que están mas cerca de x y cada uno tiene un peso vecindad.
- 3.) Sea $\lambda_q(x)$ la distancia del q -ésimo valor x_i mas lejano de x . Sea también W la función de peso tricúbica:

$$W(u) = \begin{cases} (1-u^3)^3, & \text{para } 0 \leq u < 1 \\ 0, & \text{en otro caso.} \end{cases}$$

- 4.) Se define el peso vecindad para cualquier x_i como

$$v_i(x) = W\left(\frac{|x_i - x|}{\lambda_q(x)}\right),$$

teniendo un alto peso los valores x_i cercanos a x .

- 5.) Finalmente se ajusta un polinomio de grado d a los datos, con pesos $v_i(x)$ en (x_i, y_i) . El valor del polinomio localmente ajustado en x es $\hat{g}(x)$.

Los archivos `descomposicion.R` y `descomposicion.ipynb` tiene la forma de implementar la metodología anterior en R y Python. Podemos utilizar el libro de web <https://otexts.com/fpp2/decomposition.html> para el uso de software en R.

Note que en principio, la utilidad de la descomposición en series de tiempo es exploratoria, sin embargo, dependiendo de las descomposición, ésta también

puede ser usada para obtener un pronóstico. Para llevar a cabo la predicción, vamos a asumir que la serie descompuesta puede ser escrita como

$$y_t = \hat{S}_t + \hat{A}_t$$

donde \hat{S}_t es la componente estacional, $\hat{A}_t = \hat{T}_t + \hat{R}_t$ es la componente de ajustada estacionalmente. La idea consistirá en hacer el pronóstico de las componentes \hat{S}_t y \hat{A}_t separadamente. Vamos a asumir que la componente estacional no cambia a lo largo del tiempo o que cambia de forma lenta, tal que, su pronóstico es simplemente el último ciclo estacional, es decir, el método ingenuo(naive). Para hacer el pronóstico de la componente \hat{A}_t , cualquier método de pronóstico no estacional puede ser usado. Por ejemplo, usando un modelo de una caminata aleatoria con drfit o usando un modelo ARIMA no estacional. Para mas detalles, ir a la página recomendada.

7.2 Suavizamiento Exponencial

Sea a_t la estimación de m_t , consideremos entonces

$$a_t = \alpha x_t + (1 - \alpha)a_{t-1} \quad 0 < \alpha < 1.$$

a_t : es un promedio ponderado de nuestras observaciones desde el tiempo t. También es conocido como EWMA(exponentially weighted moving average).
 α : Parámetro de suavizado. Cuando $\alpha \rightarrow 1$, la componente de tendencia a_t es poco suave y se aproxima a x_t .

Se puede verificar iterando que:

$$a_t = \alpha x_t + \alpha(1 - \alpha)x_{t-1} + \alpha(1 - \alpha)^2x_{t-2} + \cdots,$$

note que este es un caso especial de un filtro. Ahora,

$$\hat{m}_t = \alpha x_t + (1 - \alpha)\hat{m}_{t-1}, \quad \hat{m}_1 = x_1.$$

Note que

$$\hat{x}_{t+k|t} = a_t, \quad k = 1, 2, \dots,$$

La forma de estimar α es minimizando los errores de predicción un paso adelante

$$SS1RE = \sum_{t=2}^n e_t^2$$

con $e_t = x_t - \hat{x}_{t|t-1} = x_t - a_{t-1}$. Note que el valor de α es mas un hiper-parámetro que un parámetro en si.

La idea de Holt-Winters fue la de usar los promedios móviles ponderados exponencialmente para actualizar las estimaciones de la media ajustada estacionalmente(nivel), pendiente(cambio de nivel de un periodo a otros) y estaciones.

Consideremos que para una serie $\{x_t\}$ con periodo p , la actualización de las ecuaciones son:

$$\begin{aligned} a_t &= \alpha(x_t - S_{t-p}) + (1 - \alpha)(a_{t-1} + b_{t-1}), \\ b_t &= \beta(a_t - a_{t-1}) + (1 - \beta)b_{t-1}, \\ s_t &= \gamma(x_t - a_t) + (1 - \gamma)S_{t-p}, \end{aligned}$$

a_t : nivel

b_t : pendiente

s_t : efecto estacional, donde α, β, γ son los parámetros de suavizamiento y $\hat{x}_{n+k|n} = a_n + kb_n + S_{n+k-p}$. La forma de estimar los parámetros de suavizamiento es otra vez minimizando los errores de predicción un paso adelante al cuadrado(suma de los residuales al cuadrado), pero ésta vez la predicción un paso adelante es obtenida a través de $\hat{x}_{n+k|n}$. Existe muchas variantes del método de suavizamiento exponencial el cual fue debido originalmente a Holt-Winters. En algunos casos se puede incluir directamente la transformación de Box-Cox.

Nota 7.2. *El método de suavizamiento exponencial puede modificarse para incluir una constante de amortiguamiento que evita sobre-pronosticar en horizontes largos de pronóstico, es decir, trata de aplandar un poco las tendencias y la estacionalidad, ver libro fpp3 sección 8.4.*

Table 8.6: Formulas for recursive calculations and point forecasts. In each case, ℓ_t denotes the series level at time t , b_t denotes the slope at time t , s_t denotes the seasonal component of the series at time t , and m denotes the number of seasons in a year; α , β^* , γ and ϕ are smoothing parameters, $\phi_h = \phi + \phi^2 + \dots + \phi^h$, and k is the integer part of $(h - 1)/m$.

Trend	N	Seasonal		M
		A		
N	$\hat{y}_{t+h t} = \ell_t$	$\hat{y}_{t+h t} = \ell_t + s_{t+h-m(k+1)}$		$\hat{y}_{t+h t} = \ell_t s_{t+h-m(k+1)}$
	$\ell_t = \alpha y_t + (1 - \alpha)\ell_{t-1}$	$\ell_t = \alpha(y_t - s_{t-m}) + (1 - \alpha)\ell_{t-1}$		$\ell_t = \alpha(y_t/s_{t-m}) + (1 - \alpha)\ell_{t-1}$
	$s_t = \gamma(y_t - \ell_{t-1}) + (1 - \gamma)s_{t-m}$			$s_t = \gamma(y_t/\ell_{t-1}) + (1 - \gamma)s_{t-m}$
A	$\hat{y}_{t+h t} = \ell_t + hb_t$	$\hat{y}_{t+h t} = \ell_t + hb_t + s_{t+h-m(k+1)}$		$\hat{y}_{t+h t} = (\ell_t + hb_t)s_{t+h-m(k+1)}$
	$\ell_t = \alpha y_t + (1 - \alpha)(\ell_{t-1} + b_{t-1})$	$\ell_t = \alpha(y_t - s_{t-m}) + (1 - \alpha)(\ell_{t-1} + b_{t-1})$		$\ell_t = \alpha(y_t/s_{t-m}) + (1 - \alpha)(\ell_{t-1} + b_{t-1})$
	$b_t = \beta^*(\ell_t - \ell_{t-1}) + (1 - \beta^*)b_{t-1}$	$b_t = \beta^*(\ell_t - \ell_{t-1}) + (1 - \beta^*)b_{t-1}$		$b_t = \beta^*(\ell_t - \ell_{t-1}) + (1 - \beta^*)b_{t-1}$
A_d	$\hat{y}_{t+h t} = \ell_t + \phi_h b_t$	$\hat{y}_{t+h t} = \ell_t + \phi_h b_t + s_{t+h-m(k+1)}$		$\hat{y}_{t+h t} = (\ell_t + \phi_h b_t)s_{t+h-m(k+1)}$
	$\ell_t = \alpha y_t + (1 - \alpha)(\ell_{t-1} + \phi b_{t-1})$	$\ell_t = \alpha(y_t - s_{t-m}) + (1 - \alpha)(\ell_{t-1} + \phi b_{t-1})$		$\ell_t = \alpha(y_t/s_{t-m}) + (1 - \alpha)(\ell_{t-1} + \phi b_{t-1})$
	$b_t = \beta^*(\ell_t - \ell_{t-1}) + (1 - \beta^*)\phi b_{t-1}$	$b_t = \beta^*(\ell_t - \ell_{t-1}) + (1 - \beta^*)\phi b_{t-1}$		$b_t = \beta^*(\ell_t - \ell_{t-1}) + (1 - \beta^*)\phi b_{t-1}$
	$s_t = \gamma(y_t - \ell_{t-1} - \phi b_{t-1}) + (1 - \gamma)s_{t-m}$	$s_t = \gamma(y_t/(ell_{t-1} + \phi b_{t-1})) + (1 - \gamma)s_{t-m}$		$s_t = \gamma(y_t/(ell_{t-1} + \phi b_{t-1})) + (1 - \gamma)s_{t-m}$

Figura 7.1:

7.3 Rolling y Evaluación de los pronósticos

Antes de hablar de la evaluación de los pronósticos, hablaremos que hay dos tiempos de pronósticos, como los define [19] en el libro fpp3:

- ✓ Los pronósticos ex-ante son aquellos que se realizan utilizando únicamente la información que está disponible de antemano. Estos son pronósticos genuinos, hechos con anticipación utilizando cualquier información disponible en ese momento.
- ✓ Las previsiones ex-post son aquellas que se realizan utilizando información posterior sobre los predictores. Estos no son pronósticos genuinos, pero son útiles para estudiar el comportamiento de los modelos de pronóstico.

El Rolling es un método que permite computar medidas de precisión de los pronósticos ex-ante. La idea es basada en validación cruzada para series de tiempo, en donde el origen de pronóstico va rodando. Ver ejemplo en <https://otexts.com/fpp3/tscv.html> o como sigue:

Consideremos que la serie de tiempo es de tamaño $T = 15$, y que esta es dividida en entrenamiento($l = 10$) y prueba($n = 5$) como sigue:

$$(y_1) (y_2) (y_3) (y_4) (y_5) (y_6) (y_7) (y_8) (y_9) (y_{10}) (y_{11}) (y_{12}) (y_{13}) (y_{14}) (y_{15})$$

Se desea hacer una análisis con base en los pronósticos 1 y 2-pasos adelante.

La primera ventana de Rolling es de tamaño 10 y usa los parámetros que se estimaron $\hat{\theta}_{10}$ con los datos de entrenamiento y_1, \dots, y_{10} y se producen los pronósticos 1 y 2-pasos adelante.

Primera Ventana

y_1	y_2	y_3	y_4	y_5	y_6	y_7	y_8	y_9	y_{10}	y_{11}	y_{12}
Pronóstico											
									$\hat{y}_{11 10}$	$\hat{y}_{12 10}$	
									$e_{1 1}$	$e_{1 2}$	

La segunda ventana de rolling usa los parámetros que se estimaron $\hat{\theta}_{10}$, y ahora el tamaño de la venta es de 11 que corresponde a los datos y_1, \dots, y_{11} :

Segunda Ventana

y_1	y_2	y_3	y_4	y_5	y_6	y_7	y_8	y_9	y_{10}	y_{11}	y_{12}	y_{13}
										$\hat{y}_{12 11}$	$\hat{y}_{13 11}$	
										$e_{2 1}$	$e_{2 2}$	

La tercera ventana de rolling sigue usando los parámetros que se estimaron $\hat{\theta}_{10}$, y ahora el tamaño de la venta es de 12 que corresponde a los datos y_1, \dots, y_{12} , y así sucesivamente:

Tercera Ventana

y_1	y_2	y_3	y_4	y_5	y_6	y_7	y_8	y_9	y_{10}	y_{11}	y_{12}	y_{13}	y_{14}
											$\hat{y}_{13 12}$	$\hat{y}_{14 12}$	
											$e_{3 1}$	$e_{3 2}$	

Cuarta Ventana

y_1	y_2	y_3	y_4	y_5	y_6	y_7	y_8	y_9	y_{10}	y_{11}	y_{12}	y_{13}	y_{14}	y_{15}
												$\hat{y}_{14 13}$	$\hat{y}_{15 13}$	
												$e_{4 1}$	$e_{4 2}$	

Quinta Ventana

y_1	y_2	y_3	y_4	y_5	y_6	y_7	y_8	y_9	y_{10}	y_{11}	y_{12}	y_{13}	y_{14}	y_{15}
													$\hat{y}_{15 14}$	
													$e_{5 1}$	

Nota 7.3. Existen mas variantes del rolling:

- ✓ Se puede ir estimando los parámetros del modelo cada vez que ingresa una nueva observación.

- ✓ El tamaño de la ventana puede ser fijo, es decir, se puede ir descartando las primeras observaciones para mantener el tamaño de la ventana.

7.3.1 Medidas de Precisión.

La medida mas usada es el error cuadrático medio de predicción h -pasos adelante, es decir,

$$ECM[\hat{X}_{T+h|T}] = E[(X_{T+h} - \hat{X}_{T+h|T})^2] = E[e_{T+h|T}^2],$$

donde $X_{T+h} - \hat{X}_{T+h|T} = e_{T+h|T}$ es el error de predicción h -pasos adelante y el cual estima mediante

$$\frac{1}{T - l - h + 1} \sum_{t=1}^{l-h+1} e_{t,h}^2.$$

También existen mas medidas de precisión, por ejemplo el error medio absoluto el cual se define como:

$$\frac{1}{T - l - h + 1} \sum_{t=1}^{l-h+1} |e_{t,h}|,$$

el cual es un estimador de $E[|X_{T+h} - \hat{X}_{T+h|T}|]$.

Nota 7.4. Note que el promedio es un estimador del valor esperado de interés.

7.4 Modelos Estructurales de Series de Tiempo

En la presente sección, vamos a introducir una metodología para el análisis de series de tiempo vía modelos estructurales a través de los modelos de Espacio-Estado. Por ésta razón, es necesario tener el pre-requisito de los modelos de espacio-estado y el filtro de Kalman, el libro base será [9]. Consideraremos el siguiente modelo estructural para la serie de tiempo $\{Y_t\}$

$$Y_t = \mu_t + \gamma_t + c_t + \sum_{j=1}^k \beta_j x_{jt} + \varepsilon_t$$

donde μ_t es una componente de variación lenta llamada *tendencia*, γ_t es una componente periódica de periodo fijo llamada *estacional*, c_t es la componente

cíclica de periodo mayor a la componente estacional, x_{jt} es la j -ésima variable regresora o explicativa, y ε_t es la componente irregular o de error.

Veamos una primera propuesta de modelo para cada una de las componentes. La propuesta La componente de tendencia μ_t se puede ver como una extensión dinámica del modelo de regresión de intercepto y pendiente:

$$\mu_{t+1} = \mu_t + \nu_t + \eta_{t+1}, \quad \eta_{t+1} \sim N(0, \sigma_\eta^2)$$

$$\nu_{t+1} = \nu_t + \zeta_{t+1}, \quad \zeta_{t+1} \sim N(0, \sigma_\zeta^2)$$

La componente estacional γ_t puede ser escrita como:

$$\gamma_t = -\sum_{j=1}^{s-1} \gamma_{t+1-j} + \omega_t, \quad \omega_t \sim N(0, \sigma_\omega^2)$$

La condición principal es que la la suma de las componentes estacionales en un ciclo completo es cero.

La componente cíclica trata de capturar los efectos cílicos en etapas de tiempo mas grandes que las capturadas por la componente estacional. Para el caso de ciclos económicos, ellos intentan capturar los ciclos de negocios los cuales se esperan que tengan un periodo entre 1.5 y 12 años, es decir $2\pi/\lambda_c$. La componente cíclica determinística puede ser escrita como

$$c_t = \tilde{c} \cos \lambda_c t + \tilde{c}^* \sin \lambda_c t$$

Mientras que la componente cíclica también puede considerar se de la siguiente manera con $0 < \rho_c \leq 1$:

$$c_{t+1} = \rho_c [c_t \cos \lambda_c + c_t^* \sin \lambda_c] + \tilde{\omega}_t, \quad \tilde{\omega}_t \sim N(0, \sigma_{\tilde{\omega}}^2)$$

$$c_{t+1}^* = \rho_c [-c_t \sin \lambda_c + c_t^* \cos \lambda_c] + \tilde{\omega}_t^*, \quad \tilde{\omega}_t^* \sim N(0, \sigma_{\tilde{\omega}^*}^2)$$

Se puede observar que λ_c es la frecuencia del ciclo, la cual pasa a ser un parámetro del modelo o identificado por el periodograma.

Nota:Los modelos presentados anteriormente para las diferentes componentes no son lo únicos.

El modelo de espacio y estados general lineal Gaussiano puede ser escrito como:

$$Y_t = Z_t \alpha_t + \varepsilon_t \quad \varepsilon_t \sim N(0, H_t)$$

$$\alpha_{t+1} = T_t \alpha_t + R_t \eta_t \quad \eta_t \sim N(0, Q_t)$$

Nota 7.5. La primera es la ecuación de observación y la segunda es la ecuación de estados.

Las matrices Z_t , T_t , R_t , H_t y Q_t son inicialmente conocidas y los errores ε_t y η_t son independientes para cada punto del tiempo.

En el modelo de espacio y estados se modela la tendencia, esto se puede expresar como:

$$\begin{aligned} y_t &= \mu_t + \varepsilon_t, \quad \varepsilon_t \sim N(0, \sigma_\varepsilon^2) \\ \mu_{t+1} &= \mu_t + \nu_t + \xi_t \quad \xi_t \sim N(0, \sigma_\xi^2), \\ \nu_{t+1} &= \nu_t + \zeta_t, \quad \zeta_t \sim N(0, \sigma_\zeta^2) \end{aligned}$$

Este modelo es conocido como modelo de tendencia local lineal y puede ser escrito como un modelo de espacio y estados general como:

$$\begin{aligned} y_t &= \begin{pmatrix} 1 & 0 \end{pmatrix} \begin{pmatrix} \mu_t \\ \nu_t \end{pmatrix} + \varepsilon_t \\ \begin{pmatrix} \mu_{t+1} \\ \nu_{t+1} \end{pmatrix} &= \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \mu_t \\ \nu_t \end{pmatrix} + \begin{pmatrix} \xi_t \\ \zeta_t \end{pmatrix} \end{aligned}$$

En términos generales se desea que el término de la componente estacional cambie en el tiempo, lo cual se logra sumando un término aleatorio como sigue:

$$\gamma_{t+1} = -\sum_{j=1}^{s-1} \gamma_{t+1-j} + \omega_t, \quad \omega_t \sim N(0, \sigma_\omega^2).$$

Una forma alternativa de escribir la componente estacional es mediante funciones trigonométricas es:

$$\gamma_t = \sum_{j=1}^{\lfloor s/2 \rfloor} (\tilde{\gamma}_j \cos \lambda_j t + \tilde{\gamma}_j^* \sin \lambda_j t) \quad \lambda_j = \frac{2\pi j}{s} \quad j = 1, \dots, \lfloor s/2 \rfloor$$

donde $\tilde{\gamma}$ y $\tilde{\gamma}^*$ pueden ser reemplazados por caminatas aleatorias como:

$$\tilde{\gamma}_{j,t+1} = \tilde{\gamma}_{jt} + \tilde{\omega}_{jt} \text{ y } \tilde{\gamma}_{j,t+1}^* = \tilde{\gamma}_{jt} + \tilde{\omega}_{jt}^*.$$

Teniendo en cuenta la componente estacional el modelo puede ser escrito de la siguiente forma:

$$y_t = \mu_t + \gamma_t + \varepsilon_t.$$

Si se asume la forma general del modelo las componentes estarán dadas por:

$$\begin{aligned} Z_t &= (Z_{[\mu]}, Z_{[\gamma]}) & T_t &= \text{diag}(T_{[\mu]}, T_{[\gamma]}) \\ R_t &= \text{diag}(R_{[\mu]}, R_{[\gamma]}) & Q_t &= \text{diag}(Q_{[\mu]}, Q_{[\gamma]}) \end{aligned}$$

El modelo teniendo en cuenta la componente estacional puede ser expresado como un modelo de espacio y estados general con las siguientes componentes:

$$\begin{aligned} Z_{[\mu]} &= \begin{pmatrix} 1 & 0 \end{pmatrix} & Z_{[\gamma]} &= \begin{pmatrix} 1, 0, \dots, 0 \end{pmatrix} \\ T_{[\mu]} &= \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} & T_{[\gamma]} &= \begin{pmatrix} -1 & -1 & \dots & -1 & -1 \\ 1 & 0 & & 0 & 0 \\ 0 & 1 & & 0 & 0 \\ & & \ddots & & \\ 0 & 0 & & 1 & 0 \end{pmatrix} \\ R_{[\mu]} &= I_2 & R_{[\gamma]} &= \begin{pmatrix} 1, 0, \dots, 0 \end{pmatrix}' \\ Q_{[\mu]} &= \begin{pmatrix} \sigma_\xi^2 & 0 \\ 0 & \sigma_\zeta^2 \end{pmatrix} & Q_{[\gamma]} &= \sigma_\omega^2 \end{aligned}$$

Si se asume que la componente estacional en la forma trigonométrica, las componentes del modelo de espacio y estados será:

$$\alpha_t = (\mu_t \quad \nu_t \quad \gamma_{1t} \quad \gamma_{1t}^* \quad \gamma_{2t} \quad \dots)',$$

donde el resto de componentes dependerá si el número de estaciones es par o impar.

◇ Si s es par, se tiene que $s^* = s/2$, adicionalmente:

$$\begin{aligned} Z_{[\gamma]} &= \begin{pmatrix} 1 & 0 & 1 & 0 & 1 & \dots & 1 & 0 & 1 \end{pmatrix} \\ R_{[\gamma]} &= I_{s-1} & Q_{[\gamma]} &= \sigma_\omega^2 I_{s-1} \\ c_j &= \begin{pmatrix} \cos \lambda_j & \sin \lambda_j \\ -\sin \lambda_j & \cos \lambda_j \end{pmatrix} & \lambda_j &= \frac{2\pi j}{s} \quad j = 1, \dots, s^* \\ T_{[\gamma]} &= \text{diag} \left(C_1 \quad \dots \quad c_{s^*} \quad -1 \right) \end{aligned}$$

◇ si s es impar, se tiene que $s^* = (s - 1)/2$ y:

$$\begin{aligned} Z_{[\gamma]} &= \begin{pmatrix} 1 & 0 & 1 & 0 & 1 & \dots & 1 & 0 \end{pmatrix} \\ R_{[\gamma]} &= I_{s-1} & Q_{[\gamma]} &= \sigma_\omega^2 I_{s-1} \end{aligned}$$

$$C_j = \begin{pmatrix} \cos \lambda_j & \sin \lambda_j \\ -\sin \lambda_j & \cos \lambda_j \end{pmatrix} \quad \lambda_j = \frac{2\pi j}{s} \quad j = 1, \dots, s^*$$

$$T_{[\gamma]} = \text{diag} \left(C_1 \quad \dots \quad C_{s^*} \right)$$

Ejercicio 7.6. Simule 200 observaciones del siguiente modelo de espacio y estados

$$\begin{aligned} y_t &= \mu_t + \gamma_t + \varepsilon_t & \varepsilon_t &\sim N(0, 4) \\ \mu_{t+1} &= \mu_t + \nu_t + \xi_t & \xi_t &\sim N(0, 1) \\ \nu_{t+1} &= \nu_t + \zeta_t & \zeta_t &\sim N(0, 4) \\ \gamma_{t+1} &= - \sum_{j=1}^{s-1} \gamma_{t+1-j} + \omega_t & \omega_t &\sim N(0, 9) \end{aligned}$$

Asuma que $s = 4$ y proponga los valores iniciales. Vea como es la gráfica de la serie de tiempo y_t al igual que sus componentes μ_t, ν_t y ω_t .

Otra componente importante en series de tiempo es la componente cíclica, la cual puede ser introducida extendiendo el modelo de tendencia local y estacional de la siguiente manera:

$$y_t = \mu_t + \gamma_t + c_t + \varepsilon_t,$$

en donde c_t esta dada por

$$c_t = \tilde{c} \cos \lambda_c t + \tilde{c}^* \sin \lambda_c t.$$

Es posible escribir la ecuación de la componente cíclica de manera estocástica como:

$$\begin{aligned} c_{t+1} &= c_t \cos \lambda_c + c_t^* \sin \lambda_c + \omega_t \\ c_{t+1}^* &= -c_t \sin \lambda_c + c_t^* \cos \lambda_c + \omega_t^* \end{aligned}$$

Adicionalmente se puede expresar como un modelo de espacio y estados general, en donde sus correspondientes vectores y matrices son:

$$\begin{aligned} Z_{[c]} &= \begin{pmatrix} 1 & 0 \end{pmatrix} & T_{[\gamma]} &= C_{c1} \\ R_{[c]} &= I_2 & Q_{[c]} &= \sigma_\omega^2 I_2 \end{aligned}$$

$$C_j = \begin{pmatrix} \cos \lambda_j & \sin \lambda_j \\ -\sin \lambda_j & \cos \lambda_j \end{pmatrix}$$

Nota 7.7. En general es posible expresar los modelos ARMA, ARIMA y SARIMA como un modelo de espacio y estados, lo cual vuelve más sencillo el poder estimar datos faltantes o ajustar efectos de datos atípicos vía filtro de Kalman.

7.5 Predicción de las Componentes del Modelo Estructural

Por supuesto, las componentes del modelo estructural no son observables, sin embargo, con la ayuda del filtro de Kalman podemos estimar o predecir los valores de estas componentes en cada momento del tiempo.

Filtrado

Una vez se selecciona el modelo indicado para la serie de tiempo, se procede a implementar el filtro de Kalman para hallar los valores filtrados del vector de estados, los cuales se obtienen a partir de las siguientes formas recursivas:

$$\begin{aligned} \nu_t &= y_t - Z_t a_t & F_t &= Z_t P_t Z_t' + H_t \\ a_{t|t} &= a_t + P_t Z_t' F_t^{-1} \nu_t & P_{t|t} &= P_t - P_t Z_t' F_t^{-1} Z_t P_t \\ a_{t+1} &= T_t a_t + K_t \nu_t & P_{t+1} &= T_t P_t (T_t - K_t Z_t)' + R_t Q_t R_t', \end{aligned}$$

donde $K_t = T_t P_t Z_t' F_t^{-1}$, con a_1 y P_1 son el vector de media y la matriz de varianzas del estado inicial del vector α_1 . Adicionalmente se tienen las siguientes dos relaciones.

$$a_{t+1} = T_t a_{t|t} \quad P_{t+1} = T_t P_{t|t} T_t' + R_t Q_t R_t'$$

Por notación, se ha establecido que:

$$a_{t|t} = E[\alpha_t | y_t, \dots, y_1],$$

$$P_{t|t} = \text{Var}[\alpha_t | y_t, \dots, y_1],$$

$$a_{t+1} = E[\alpha_{t+1} | y_t, \dots, y_1],$$

$$P_{t+1} = \text{Var}[\alpha_{t+1} | y_t, \dots, y_1],$$

Suavizado

Consideraremos ahora la estimación o predicción de los estados $\alpha_1, \dots, \alpha_n$ con base en la información $Y_n = (Y_1, \dots, Y_n)$, es decir, se calcula la esperanza condicional $\hat{\alpha}_t = E(\alpha_t|Y_n)$ y $V_t =$

Con ayuda de la esperanza condicional es posible definir los estados recursivos de suavizado para el caso de modelos de espacio y estados, los cuales son:

$$\begin{aligned} r_{t-1} &= Z'_t F_t^{-1} \nu_t + L'_t r_t & N_{t-1} &= Z'_t F_t^{-1} Z_t + L'_t N_t L_t \\ \hat{\alpha}_t &= a_t + P_t r_{t-1} & V_t &= P_t - P_t N_{t-1} P_t \end{aligned}$$

Las cuales corresponden a un procedimiento backward, es decir, se procede para $t = n, n-1, \dots, 1$.

Pronóstico

Sea $\bar{y}_{n+j} = E[y_{n+j}|Y_n, \dots, y_1]$, el pronóstico de y_{n+j} con base en las observaciones y_1, \dots, y_n que minimiza el error cuadrático medio, es decir $\bar{F}_{n+j} = E[(\bar{y}_{n+j} - y_{n+j})(\bar{y}_{n+j} - y_{n+j})'|y_n, \dots, y_1]$.

Análogo al caso del modelo de tendencia local, se asumirá que los pronósticos y_{n+1}, \dots, y_{n+j} son valores faltantes para así usar el filtro de Kalman, con lo cual se obtiene la siguiente formula recursiva:

$$\bar{y}_{n+j} = Z_{n+j} \bar{a}_{n+j}$$

con la matriz de error cuadrático medio

$$\bar{F}_{n+j} = Z_{n+j} \bar{P}_{n+j} Z'_{n+j} + H_{n+j}$$

donde,

$$\bar{P}_{n+j+1} = T_{n+j} \bar{P}_{n+j} T'_{n+j} + R_{n+j} Q_{n+j} R'_{n+j},$$

y

$$\bar{a}_{n+j+1} = T_{n+j} \bar{a}_{n+j}$$

para $j = 1, \dots, J$, donde T_t , R_t y Q_r , H_t son dados en la expresión del modelo de espacio y estados general.

Estimación de los Parámetros del Modelo de Espacio-Estado

La estimación de los parámetros de un modelo de espacio-estados por lo general se lleva a cabo vía máxima verosimilitud. Se asume que los parámetros de la condición inicial son conocidos, con lo cual los parámetros del modelo lo componen las matrices Z_t, T_t, H_t y Q_t , así que la verosimilitud es:

$$L(Y_n) = p(y_1, \dots, y_n) = p(y_i) \prod_{t=2}^n p(y_t|Y_{t-1}).$$

Recordemos que los errores en el modelo de espacio y estado son normales e independientes, entonces, tanto el vector de estado α_t como el vector de observación y_t son normales, así que aplicando logaritmo en ambos lados de la igualdad se tiene que:

$$\begin{aligned} \log L(Y_n) &= \sum_{t=1}^n \log p(y_t|Y_{t-1}) \\ &= -\frac{np}{2} \log 2\pi - \frac{1}{2} \sum_{t=1}^n (\log |F_t| + \nu_t' F_t^{-1} \nu_t), \end{aligned}$$

donde F_t y ν_t son calculados mediante el filtro de Kalman. Uno puede notar que en F_t y ν_t . Note que en los modelos estructurales que hemos propuesto hasta el momento, las matrices Z_t, T_t , están conformadas por componentes conocidas, así que los únicos parámetros son H_t y Q_t , pero en general no es así.

Diagnóstico

Los supuestos en los modelos de espacio y estados lineal general se pueden probar igual que en el caso de los modelos de tendencia local, es decir se comprueban sobre los residuales e_t , con:

$$e_t = \frac{\nu_t}{\sqrt{F_t}},$$

en donde, $F_t = P_t + \sigma_\varepsilon^2$ y $\nu_t = y_t - a_t$, las cuales fueron definidas en la metodología del filtro de Kalman. Los test que se tiene que llevar a cabo para validar el modelo son los usuales:

- Normalidad

- Homocedasticidad
- No autocorrelación

Variables Regresoras

El modelo general de espacio y estados puede ser escrito incluyendo variables regresoras, con estas variables es posible incorporar variables explicativas o intervenciones, obteniendo el siguiente modelo:

$$y_t = Z_t \alpha_t + X_t \beta + \varepsilon_t,$$

el cual se puede ver como un modelo de espacio y estados

$$\begin{aligned} y_t &= \begin{pmatrix} Z_t & X_t \end{pmatrix} \begin{pmatrix} \alpha_t \\ \beta_t \end{pmatrix} + \varepsilon_t \\ \begin{pmatrix} \alpha_{t+1} \\ \beta_{t+1} \end{pmatrix} &= \begin{pmatrix} T_t & 0 \\ 0 & I_k \end{pmatrix} \begin{pmatrix} \alpha_t \\ \beta_t \end{pmatrix} + \begin{pmatrix} R_t \\ 0 \end{pmatrix} \eta_t. \end{aligned}$$

Los datos faltantes pueden ser tratados de una manera natural usando el Filtro de Kalman siguiendo los lineamiento de [9].

Ejercicio 7.8. Existen varios paquetes y funciones que hacen la estimación de un modelo estructural de series de tiempo.

- ✓ **R:** *stsm* sin componente cíclica, *rucm* con componente cíclica.
- ✓ **Python:** Del módulo *statsmodels.api*, la función *UnobservedComponents* de *statsmodels.api.tsa*.

Realizar:

- ✓ Ajuste de un modelo con tendencia y estacional para la serie de pasajeros.
Producir los pronósticos.
- ✓ Ajuste de un modelo con tendencia y ciclo para la serie US real GNP.
Producir también los pronósticos.

7.6 Preliminares del Aprendizaje Automático

Estos preliminares están tomados de las notas [33] y del libro [32]. Lo primero que se va a mencionar es la diferencia entre el *Machine Learning* y *Statistical Learning*:

- ✓ El Machine Learning(aprendizaje automático) inventa modelos y algoritmos los cuales pueden ‘aprender’ de los datos de entrenamiento y están disponibles para generalizar esos hallazgos con el objeto de predecir nuevos resultados.
- ✓ El Statistical Learning(aprendizaje estadístico) es una disciplina de la estadística matemática el cual formaliza los modelos del Machine Learning y cuantifica su incertidumbre estadística. Además, los hallazgos teóricos pueden ser usados para inventar nuevos o al menos mejorar los algoritmos de Machine Learning existentes proponiendo reglas significativas para el ajuste de parámetros.

Para el caso de series de tiempo, se considerará el aprendizaje supervisado cuya configuración consiste en unos pares de datos observados (X, Y) , donde X es llamada la entrada y Y es llamada la salida. El objetivo consistirá en predecir Y a partir de X .

7.6.1 Teoría Estadística de la Decisión

Sea $(\Omega, \mathcal{A}, \mathbb{P})$ un espacio de probabilidad.

Definición 7.9. (*Configuración o entorno para el Aprendizaje Supervisado*)
Sea $\mathcal{X} \subset \mathbb{R}^d$ y $\mathcal{Y} = \{1, \dots, K\}$ o $\mathcal{Y} = \mathbb{R}$. Sea $\mathbb{P}^{(X,Y)}$ alguna distribución de probabilidad conjunta sobre $\mathcal{X} \times \mathcal{Y}$. Sea $(X_i, Y_i), i = 1, \dots, n$ variables aleatorias i.i.d con distribución $\mathbb{P}^{(X,Y)}$, estas variables aleatorias son llamadas muestras de entrenamiento o los datos de entrenamiento. Si $\mathcal{Y} = \{1, \dots, K\}$, estamos en la configuración de un problema de clasificación, mientras que si $\mathcal{Y} = \mathbb{R}$, entonces llamaremos a este un problema de regresión. Vamos a considerar que $(X, Y) \sim \mathbb{P}^{(X,Y)}$ es una realización independiente de $(X_i, Y_i), i = 1, \dots, n$.

El objetivo general del aprendizaje supervisado dentro del machine learning será entonces: Encontrar una función o aplicación $\hat{f}_n : \mathcal{X} \rightarrow \mathcal{Y}$ (computada de las muestras de entrenamiento $(X_i, Y_i), i = 1, \dots, n$), tal que la expresión $\hat{f}_n(X)$ es cercana a Y . Este objetivo se formaliza como sigue: los conjuntos \mathcal{X}, \mathcal{Y} son equipados con las σ -álgebras $\mathcal{B}(\mathcal{X})\mathcal{B}(\mathcal{Y})$, tal que la medibilidad de las variables X, Y está bien definida.

Definición 7.10. • Una regla de decisión es una función medible $f : \mathcal{X} \rightarrow \mathcal{Y}$

- Una función de pérdida es una función medible $L : \mathcal{Y} \times \mathcal{Y} \rightarrow \mathbb{R}^+ \cup 0$. La aplicación de una regla de decisión f a algún X produce la pérdida $L(Y, f(X))$.
- El riesgo de una regla de decisión f se define por $R(f) = \mathbb{E}L(Y, f(X))$.

Definición 7.11. Una aplicación $f^* : \mathcal{X} \rightarrow \mathcal{Y}$ el cual satisface $R(f^*) = \min_{g:\mathcal{X} \rightarrow \mathcal{Y} \text{ medible}} R(g)$ es llamado la regla de Bayes. El riesgo correspondiente $R(f^*)$ es llamado el riesgo de Bayes.

Nota 7.12. Es importante ver que para algunas específicas funciones de pérdida, hay formas explícitas para f^* . Por ejemplo (si $\mathbb{P}^{(X,Y)}$ es conocida):

- Bajo la perdida cuadrática $L(x, y) = (x - y)^2$, entonces en el problema de regresión,

$$R(f) = \mathbb{E}L(Y, f(X)) = \mathbb{E}[(Y - f(X))^2] = \int \int (y - f(x))^2 d\mathbb{P}^{Y|X=x}(y) d\mathbb{P}^X(x)$$

la cual se hace mínimo para $f^* = \mathbb{E}[Y|X = x]$, es decir la esperanza condicional de Y dado $X = x$.

- Para el problema de clasificación, basados en la función de pérdida 0-1, es decir, toma el valor de 0 si la decisión es correcta, y toma el valor de 1 si la decisión es incorrecta. Con esto, el riesgo queda de la siguiente manera:

$$R(f) = \mathbb{E}L(Y, f(X)) = E[\mathbb{1}_{\{Y \neq f(X)\}}] = \mathbb{P}(Y \neq f(X)) = \int \mathbb{P}(Y \neq f(X)|X = x) d\mathbb{P}^X(x)$$

, la cual es mínima para

$$f^*(x) = \arg \min_{k \in 1, \dots, K} \mathbb{P}(Y \neq k | X = x) = \arg \max_{k \in 1, \dots, K} \mathbb{P}(Y = k | X = x).$$

Este resultado se entiende que f^* selecciona la clase que es más probable dada una observación x , en este caso f^* es llamado el clasificador MAP (máximo a posteriori). Al igual que $\mathbb{P}^{(X,Y)}$, f^* se desconoce en las aplicaciones, pero el objetivo es, por supuesto, determinar una regla de decisión que sea lo más “cercana” posible a f^* . Llamamos algoritmo al procedimiento para determinar una regla de decisión basada en datos de entrenamiento.

Nota 7.13. Note que esta formulación general es análoga a la propuesta en [40], la cual radica en lo siguiente basados en la figura 7.2:

El modelo general de aprendizaje consiste de:

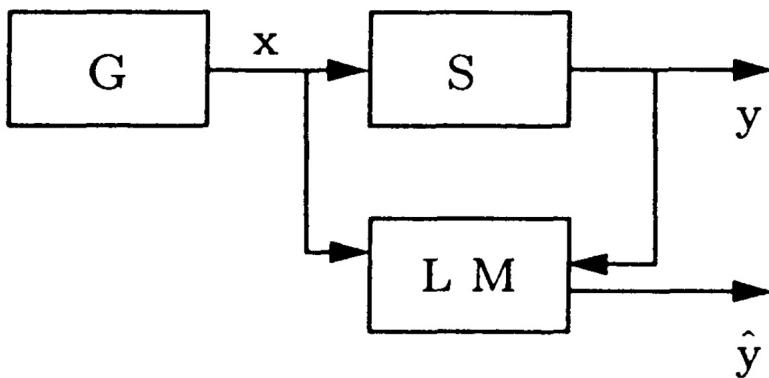


Figura 7.2:

- i) Un generador(G) de vectores aleatorios $X \in \mathbb{R}^p$, que se extraen independientemente de distribución de probabilidad fija $F_X(x)$ pero desconocida.
- ii) Un supervisor(S) el cual retorna un valor de salida $Y = y$ a cada entrada $X = x$, de acuerdo a la distribución condicional $F_{Y|X=x}(y|x)$, la cual también es fija pero desconocida.
- iii) Una máquina de aprendizaje(LM) capaz de implementar un conjunto de funciones $f(x, \alpha), \alpha \in \Lambda$, donde Λ es el espacio de parámetros.

El problema de aprendizaje es aquel que consiste en elegir de un conjunto de funciones dadas $f(x, \alpha), \alpha \in \Lambda$, aquella que mejor aproxima la respuesta del supervisor, basados en una función de pérdida o discrepancia, $L(y, f(x, \alpha))$, o más específicamente sobre el valor esperado de la función de pérdida, la cual llamaremos riesgo funcional y que se define como

$$R(\alpha) = \int L(y, f(x, \alpha))dF(x, y).$$

Pos supuesto la elección de la función $f(x, \alpha), \alpha \in \Lambda$ se debe basa en una muestra llamada conjunto de entrenamiento, de n observaciones i.i.d extraídas de la distribución $F(x, y) = F(x)F(y|x)$:

$$(x_1, y_1), \dots, (x_n, y_n).$$

El problema en forma general es como sigue: Considere la medida de probabilidad $F(z)$ definida sobre el espacio Z . Considere el conjunto de funciones $Q(z, \alpha), \alpha \in \Lambda$. El objetivo es minimizar el riesgo funcional

$$R(\alpha) = \int Q(z, \alpha) dF(z), \alpha \in \Lambda,$$

donde la medida de probabilidad $F(z)$ es desconocida, pero una muestra i.i.d z_1, \dots, z_n está dada. Note que z describe los pares (x, y) y $Q(z, \alpha)$ es una función de pérdida específica y α no necesariamente es un vector en forma general.

7.6.2 Evaluación de Algoritmos

Dada los datos de entrenamiento $(X_i, Y_i), i = 1, \dots, n$, el objetivo es definir una regla de decisión \hat{f}_n la cual esté lo mas cercano posible a f^* . La función \hat{f}_n el cual mapea las muestras de entrenamiento a la regla de decisión es llamado algoritmo.

Definición 7.14. (*Algoritmo*) Una función o aplicación medible $\hat{f}_n : \Omega \times \mathcal{X} \rightarrow \mathcal{Y}$ es llamado un algoritmo, si

- (i) Para cada $x \in \mathcal{X}$ fijo, la función $\hat{f}_n(\omega, x) : \Omega \rightarrow \mathcal{Y}$ es medible con respecto a $T_n = ((X_i, Y_i)), i = 1, \dots, n$. Mas específicamente, si existe una aplicación $A : (\mathcal{X} \times \mathcal{Y})^n \times \mathcal{X} \rightarrow \mathcal{Y}$, tal que para todo $\omega \in \Omega, x \in \mathcal{X}$, , se satisface que

$$\hat{f}_n(\omega, x) = A(T_n(\omega), x).$$

- (ii) Para cada $\omega \in \Omega$ fijo $\hat{f}_n(\omega, x) : \mathcal{X} \rightarrow \mathcal{Y}$ es una regla de decisión.

Nota 7.15. Con abuso de esta terminología, también se le llamará regla de decisión a \hat{f}_n en el sentido del numeral (ii) de la definición de algoritmo.

Definición 7.16. (*Error de Generalización*)

$\mathbb{E}R(\hat{f}_n)$ se llama error de generalización, donde

$$R(\hat{f}_n) := \mathbb{E}[L(Y, \hat{f}_n(X)) | T_n]$$

con $(X, Y) \sim \mathbb{P}^{(X, Y)}$ independiente de T_n .

Nota 7.17. Desde un punto de vista teórico, la evaluación de la calidad de \hat{f}_n se realiza en dos pasos:

- La esperanza condicional $R(\hat{f}_n) = \mathbb{E}[L(Y, \hat{f}_n(X))|T_n]$ calcula el riesgo de regla de decisión \hat{f}_n inducida bajo el supuesto de que la observación utilizada para la evaluación (X, Y) es independiente de los datos de entrenamiento. Sin embargo, $R(\hat{f}_n)$ no evalúa los datos de entrenamiento que se utilizaron para determinar a \hat{f}_n . Por ejemplo, puede ser que los datos de entrenamiento T_n no fueran representativos de la relación general entre X e Y debido a coincidencias "desfavorables". En este caso, $R(\hat{f}_n)$ es grande, aunque el algoritmo \hat{f}_n puede ser muy bueno.
- Solo el error de generalización $\mathbb{E}R(\hat{f}_n)$ evalúa el algoritmo \hat{f}_n , es decir, la regla con la cual las reglas de decisión se obtienen de T_n promediando el riesgo $R(\hat{f}_n)$ sobre todas las configuraciones de datos de entrenamiento teóricamente posibles.

Para los resultados teóricos y la derivación de nuevos algoritmos, se emplea sobre todo $\mathbb{E}R(\hat{f}_n)$ o estadísticas sobre el tamaño de $R(\hat{f}_n)$ son de interés, mientras que en la práctica, la determinación de $R(\hat{f}_n)$ para la regla de decisión $\hat{f}(\omega)$ obtenida específicamente, es de interés, es decir, una versión muestral del riesgo.

Definición 7.18. (*Exceso de Riesgo de Bayes*)

- $\mathbb{E}R(\hat{f}_n) - R(f^*) \geq 0$ se denomina *Exceso de Riesgo de Bayes*. La expresión indica cuán fuertemente se excede el riesgo Bayesiano por el error de generalización.
- Para cada $d \in \mathbb{N}$ sea $(\psi_d(n))_{n \in \mathbb{N}}$ una secuencia decreciente monótona. Un algoritmo aprende en promedio con la tasa de convergencia $\psi_d(n)$, si existe una constante $C > 0$ independiente de d , n , de modo que

$$\text{Para todo } n \in \mathbb{N} : \mathbb{E}R(\hat{f}_n) - R(f^*) \leq C \cdot \psi_d(n). \quad (7.3)$$

Un algoritmo aprende con alta probabilidad con tasa de convergencia $\psi_d(n)$, si

$$\lim_{c \rightarrow \infty} \sup_{d,n} \lim_{d,n \rightarrow \infty} \sup \mathbb{P} \left(|R(\hat{f}_n) - R(f^*)| \geq c \cdot \psi_d(n) \right) = 0 \quad (7.4)$$

Dados los datos de entrenamiento (X_i, Y_i) , $i = 1, \dots, n$, entonces un requisito obvio para \hat{f}_n parece ser:

$$\text{para todo } i \in \{1, \dots, n\} : L \left(\hat{f}_n(X_i), Y_i \right) = 0 \quad (7.5)$$

Para muchas funciones de pérdida esto equivale a:

$$\text{para todo } i \in \{1, \dots, n\} : \hat{f}_n(X_i) = Y_i.$$

es decir, la reproducción correcta de los datos de entrenamiento. Sin embargo este enfoque tiene 2 desventajas:

- No se reconoce que X_i, Y_i están distorsionados por influencias aleatorias del entorno y que algunos Y_i no corresponden en absoluto a las "mejores respuestas posibles" $f^*(X_i)$. En el peor caso, hay dos observaciones (X_{i1}, Y_{i1}) y (X_{i2}, Y_{i2}) con $X_{i1} = X_{i2}$ pero $Y_{i1} \neq Y_{i2}$ tal que el cumplimiento de la Ecuación (7.5) no es posible en absoluto. Si Ecuación (7.5) se cumple completamente o al menos de forma aproximada para muchos índices i , hablamos de un *sobreajuste a los datos de entrenamiento*.
- La Ecuación (7.5) no da información sobre cómo elegir $\hat{f}_n(x)$ para $x \notin \{X_1, \dots, X_n\}$. En la práctica, sin embargo, \hat{f}_n también debería aplicarse a casos que no se han visto antes.

Para corregir y expandir el enfoque de la Ecuación (7.5) es necesaria una suposición sobre la estructura de f^* . Esto debe permitirnos estimar el valor de $f^*(x)$ y también para $x \in \mathcal{X}$ en un entorno de datos de entrenamiento $X_i, i = 1, \dots, n$. Normalmente, la formulación de una hipótesis sobre f^* se hace especificando una clase de funciones:

Definición 7.19. Si $\mathcal{F} \subset \{f : \mathcal{X} \rightarrow \mathcal{Y} \text{ medible}\}$ es un conjunto arbitrario de funciones, llamamos a $f^* \in \mathcal{F}$ un *Modelo supuesto sobre f^** .

Un algoritmo asociado \hat{f}_n generalmente solo selecciona elementos de \mathcal{F} .

A menudo, el modelo supuesto es una simplificación de la relación real entre X e Y, y no se aplica en la realidad (es decir $f^* \notin \mathcal{F}$). En este caso, se aplica la siguiente descomposición del exceso de riesgo de Bayes:

$$\mathbb{E}R(\hat{f}_n) - R(f^*) = \underbrace{\left[\mathbb{E}R(\hat{f}_n) - \inf_{f \in \mathcal{F}} R(f) \right]}_{\text{Error de estimación}} + \underbrace{\left[\inf_{f \in \mathcal{F}} R(f) - R(f^*) \right]}_{\text{Error de aproximación}}$$

El *error de aproximación* es entonces un error inevitable debido a nuestra simplificación. Por lo tanto, para determinar un buen algoritmo, hay que

minimizar el error de estimación.

El siguiente problema de regresión simple pretende ilustrar los conceptos con un ejemplo:

Ejemplo 7.20. Sean $X : \Omega \rightarrow \mathcal{X} = \mathbb{R}$, y aplica ("en la realidad")

$$Y = f_0(X) + \epsilon,$$

con $f_0 : \mathbb{R} \rightarrow \mathbb{R}$ una función *discontinua* y su error aleatorio ϵ . Sea $\mathbb{E}\epsilon = 0$, y que ϵ sea independiente de X .

El contexto real nos es desconocido. Hacemos el *modelo supuesto*:

$$Y = f_c(X) + \epsilon$$

con una función *continua* $f : \mathbb{R} \rightarrow \mathbb{R}$.

Utilizando la función de pérdida cuadrática $L(y, s) = (y - s)^2$, entonces se tiene que:

$$f^*(x) = \mathbb{E}[Y|X = x] = f_0(x),$$

es decir, el supuesto de nuestro modelo es $f_0 = f^* \in \mathcal{F} := \{f : \mathbb{R} \rightarrow \mathbb{R}\text{ continua}\}$. Ahora el riesgo Bayesiano es

$$R(f^*) = \mathbb{E}L(Y, f^*(X)) = \mathbb{E}[\epsilon^2]$$

Ahora, el riesgo para cualquier función de decisión $f \in \mathcal{F}$ es:

$$R(f) = \mathbb{E} \left[\left(\underbrace{Y}_{=f_0(x)+\epsilon} - f(X) \right)^2 \right] = \mathbb{E}[(f_0(X) - f(X))^2] + \mathbb{E}[\epsilon^2].$$

es decir, el *error de aproximación* causado por la suposición del modelo es

$$\inf_{f \in \mathcal{F}} R(f) - E(f^*) = \inf_{f \in \mathcal{F}} \mathbb{E}[(f_0(X) - f(X))^2].$$

El tamaño del error de aproximación depende de la frecuencia con la que ocurren las observaciones X cerca de los puntos de discontinuidad de f_0 (donde el supuesto de nuestro modelo causa los errores más grandes).

Al elegir la hipótesis del modelo $\mathcal{F} \subset \{f : \mathcal{X} \rightarrow \mathcal{Y} \text{ medible}\}$ hay un dilema natural: Dado que \hat{f}_n debería ofrecer la mejor función de decisión posible de todos los \mathcal{F} , pero solo hay n datos de entrenamiento disponibles para esto, la calidad de \hat{f}_n depende del tamaño de \mathcal{F} :

- ★ Si \mathcal{F} es grande, el error de aproximación es pequeño, pero el error de estimación es grande.
- ★ Si \mathcal{F} es pequeño, el error de aproximación es grande, pero el error de estimación pequeño.

En muchos modelos se especifica un elemento de \mathcal{F} especificando parámetros, es decir, $\mathcal{F} = \{f_\theta : \theta \in \Theta\}$, donde Θ es un espacio de parámetros. En el aprendizaje automático, a menudo se aplica $\Theta \subset \mathbb{R}^p$ con dimensión $p \in \mathbb{N}$. El término *complejidad* de \mathcal{F} se refiere vagamente al valor de $\frac{p}{d}$ (*¿cuántos parámetros se utilizan para describir un componente?*) o hasta qué punto \mathcal{F} todavía está lejos de $\{f : \mathcal{X} \rightarrow \mathcal{Y} \text{ mensurable}\}$

El estudio del error de aproximación suele ser una tarea puramente analítica y no es nuestro principal objetivo. Por lo tanto, a menudo asumiremos que

$$f^* \in \mathcal{F} \text{ y por lo tanto } \inf_{f \in \mathcal{F}} R(f) - R(f^*) = 0.$$

Para cada método de aprendizaje supervisado presentado en este libro, indicamos claramente la hipótesis del modelo $f^* \in \mathcal{F}$, de modo que una evaluación del error de aproximación también es posible en problemas de aplicación específicos.

Si existe un problema de regresión, hay una descomposición adicional del Exceso de Riesgo de Bayes independiente de la discusión anterior:

Lema 7.21. (*Exceso de riesgo de Bayes para la función de pérdida cuadrática*)
Si $\mathcal{Y} \subset \mathbb{R}$, L es la función de pérdida cuadrática y \hat{f}_n es un algoritmo arbitrario, entonces

$$\mathbb{E}R(\hat{f}_n) - R(f^*) = \mathbb{E} \text{MSE}\left(\hat{f}_n(X)\right) = \mathbb{E}\left[\mathbb{E}\left[(\hat{f}_n(X) - f^*(X))^2 | X\right]\right]$$

donde

$$\text{MSE}(\hat{f}_n(x)) := \mathbb{E}[(\hat{f}_n(x) - f^*(x))^2], \quad x \in \mathcal{X}$$

es el error cuadrático medio del algoritmo \hat{f}_n .

Se aplica la denominada descomposición de la varianza-sesgo

$$MSE(\hat{f}_n(x)) = Var(\hat{f}_n(x)) + \left| Bias(\hat{f}_n(x)) \right|^2, \quad (7.6)$$

donde

- ★ $Bias(\hat{f}_n(x)) := \mathbb{E}\hat{f}_n(x) - f^*(x)$ se llama el sesgo de $\hat{f}_n(x)$ y
- ★ $Var(\hat{f}_n(x)) = \mathbb{E} \left[(\hat{f}_n(x) - \mathbb{E}\hat{f}_n(x))^2 \right]$ la varianza de $\hat{f}_n(x)$.

Prueba Si L es la función de pérdida cuadrática, entonces $f^*(X) = \mathbb{E}[Y|X]$ se mantiene (ver Lema 1.3) y por lo tanto:

$$\begin{aligned} \mathbb{E}R(\hat{f}_n(x)) &= \mathbb{E} \left[(Y - f^*(X) + f^*(X) - \hat{f}_n(X))^2 \right] \\ &= \mathbb{E} \left[(Y - f^*(X))^2 \right] + 2 \underbrace{\mathbb{E} \left[\mathbb{E} \left[(Y - f^*(X)) \cdot (f^*(X) - \hat{f}_n(X)) | X, T \right] \right]}_{=\mathbb{E}[f^*(X)-\hat{f}_n(X)\cdot\mathbb{E}[(Y-f^*(X))|X]]=0} \\ &\quad + \mathbb{E} \left[(f^*(X) - \hat{f}_n(X))^2 \right] \\ &= R(f^*) + \mathbb{E} \left[\mathbb{E} \left[(\hat{f}_n(X) - f^*(X))^2 | X \right] \right] \end{aligned} \quad (7.7)$$

El cálculo de los dos términos $\mathbb{E}R(\hat{f}_n(x))$ y $R(f^*)$ puede así trasladarse a la investigación de cómo se comporta la diferencia $\hat{f}_n(x) - f^*(x)$ para todo $x \in \mathcal{X}$. La ecuación (7.6) evalúa el algoritmo sobre la base de dos características:

- ✓ La varianza mide la intensidad con la que el algoritmo depende de los datos de entrenamiento, es decir, la intensidad con la que un cambio en los puntos de datos de entrenamiento individuales afecta a la regla de decisión inducida por el algoritmo.
- ✓ El sesgo describe el error sistemático medio o esperado de las reglas de decisión generadas por $\hat{f}_n(x)$. Un error sistemático surge cuando el algoritmo favorece ciertas reglas de decisión en \mathcal{F} introduciendo adicionalmente restricciones. Esto puede hacerse penalizando las propiedades indeseables de $\hat{f}_n(x)$ o restringiendo la búsqueda de $\hat{f}_n(x)$ a una clase más pequeña de funciones $\hat{\mathcal{F}} \subset \mathcal{F}$ (aunque solo se sabe o se supone que $f^* \in \mathcal{F}$).

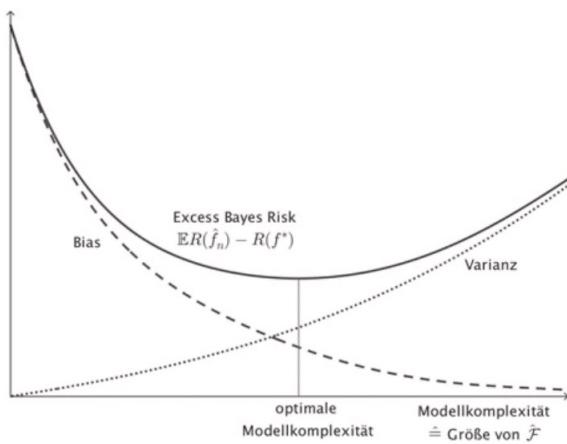


Figura 7.3: Gráfico Representación de exceso de Bayes Riesgos $\mathbb{E}R(\hat{f}_n) - R(f^*)$ y de los dos sumandos varianza y sesgo dependiendo del tamaño de la clase de función $\hat{\mathcal{F}} \subset \mathcal{F}$

Inicialmente, no parece sensato introducir deliberadamente un error sistemático a través de las posibilidades anteriormente mencionadas. Sin embargo, se ha demostrado que un error sistemático moderado puede suponer un ahorro considerable de varianza. De hecho, un sesgo pequeño se asocia a menudo con una varianza alta y un sesgo grande con una varianza pequeña; esto se denomina *compensación sesgo-varianza*. En este sentido, los algoritmos con un sesgo muy pequeño no son necesariamente los que tienen el menor error de generalización. La figura 7.3 muestra un curso típico del exceso de riesgo de Bayes $\mathbb{E}R(\hat{f}_n) - R(f^*)$ dependiendo del tamaño de la clase de función $\hat{\mathcal{F}} \subset \mathcal{F}$.

7.6.3 Enfoques estándar para la determinación de algoritmos

En esta sección presentamos un enfoque sistemático para la determinación del algoritmo. Asumimos que un modelo supuesto $f^* \in \mathcal{F}$ con la clase de función apropiada $\mathcal{F} \subset \{f : \mathcal{X} \rightarrow \mathcal{Y} \text{ medible}\}$ se ha formulado y existe una función de pérdida L .

Entonces idealmente también debe cumplirse que $\hat{f}_n \in \mathcal{F}$ y $R(\hat{f}_n)$ debe

Ejemplo Sobreajuste

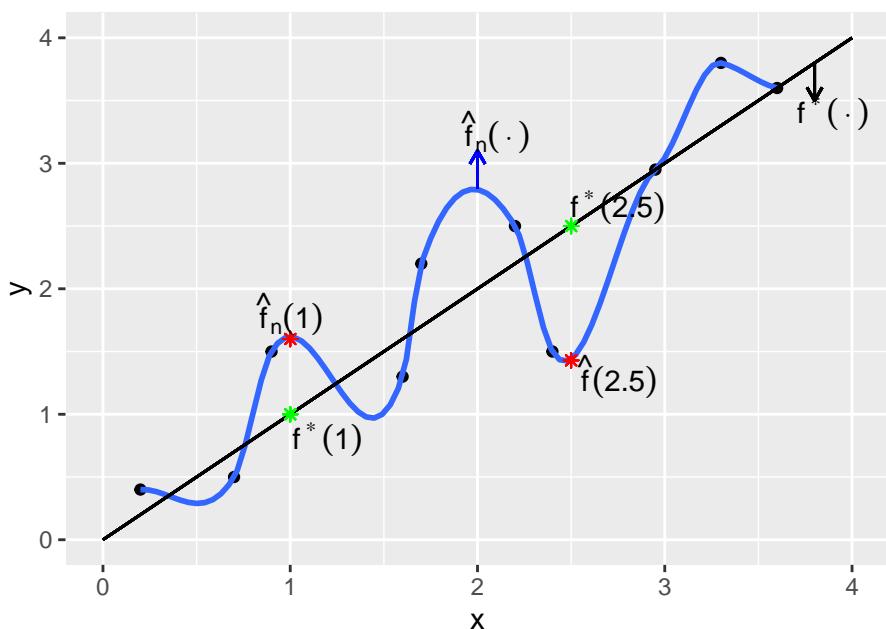


Figura 7.4: El sobreajuste o exceso de ajuste en un ejemplo de regresión lineal simple.

ser lo más pequeño posible; idealmente debería

$$\hat{f}_n \in \arg \min_{f \in \mathcal{F}} R(f) = \{f \in \mathcal{F} : R(g) \geq R(f), \text{ para todo } g \in \mathcal{F}\} \quad (7.10)$$

Para cualquier regla de decisión f , el riesgo $R(f)$ se puede estimar de la siguiente manera:

Definición 7.22. Sean los datos de entrenamiento (X_i, Y_i) , $i = 1, \dots, n$ y sea $f : \mathcal{X} \rightarrow \mathcal{Y}$ una regla de decisión, entonces

$$\hat{R}_n(f) := \frac{1}{n} \sum_{i=1}^n L(Y_i, f(X_i))$$

se llama el riesgo empírico en los datos de entrenamiento.

Si reemplazamos $R(f)$ en la ecuación (7.10) por $\hat{R}_n(f)$, obtenemos un enfoque estándar para determinar algoritmos.

Nota 7.23. (Enfoque para la determinación de algoritmos 1)

$$\hat{f}_n \in \arg \min_{f \in \mathcal{F}} \hat{R}_n(f), \quad \hat{R}_n(f) = \frac{1}{n} \sum_{i=1}^n L(Y_i, f(X_i)).$$

Si los elementos de \mathcal{F} dependen de un número muy grande o incluso infinito de parámetros libremente seleccionables, entonces llamamos a \mathcal{F} de alta dimensión o de alta complejidad. En este caso, el problema de minimización anterior no se puede resolver sin ambigüedades y una solución \hat{f}_n está sobreajustada a los datos de entrenamiento (Ecuación (7.5)). Al castigar las propiedades indeseables de \hat{f}_n , se puede reducir el número de posibles soluciones. Para esto, introducimos un término de castigo $J : \mathcal{F} \rightarrow \mathbb{R}$ (engl. penalty term), que tiene valores pequeños para $f \in \mathcal{F}$ deseables y valores grandes para rendimientos de $f \in \mathcal{F}$ no deseados. J se agrega al problema de minimización con una suma adicional:

Nota 7.24. (Enfoque para la determinación de algoritmos 2)

$$\hat{f}_{n,\lambda} \in \arg \min_{f \in \mathcal{F}} \{\hat{R}_n(f) + \lambda \cdot J(f)\} \quad (7.11)$$

donde $\lambda \geq 0$ se denomina parámetro de complejidad o parámetro de penalización.

λ indica que tan fuerte es la penalización $J(f)$ de f en comparación con $\hat{R}_n(f)$. En el caso de \mathcal{F} de alta dimensión, es necesaria una buena elección de λ para obtener un algoritmo \hat{f}_n con un error de generalización bajo.

Nota 7.25. (*Conexión entre el término de castigo y el sesgo*)

Bajo supuestos de convexidad en $\hat{R}_n(f)$ y $J(f)$ en f se puede mostrar que (ver Definición 2.32 [32]): la Ec.(7.11) es equivalente a que $\hat{f}_{n,\lambda}$ sea la solución de problema de minimización

$$\hat{f}_{n,\lambda} \in \arg \min_{f \in \hat{\mathcal{F}}_\lambda} \hat{R}_n(f)$$

donde $\hat{\mathcal{F}}_\lambda = \{f \in \mathcal{F} : J(f) \leq \hat{t}(\lambda)\}$ y $\hat{t}(\lambda)$ es un número real que depende de λ y los datos de entrenamiento (X_i, Y_i) , $i = 1, \dots, n$. Por lo general $\lambda \mapsto \hat{t}(\lambda)$ es monotonamente decreciente, esto significa que para algún λ creciente hay una reducción sistemática del espacio $\hat{\mathcal{F}}_\lambda$. Si se aplica $\hat{f}_n \notin \hat{\mathcal{F}}_\lambda$, $\hat{f}_{n,\lambda}$ da como resultado una falsificación sistemática del algoritmo basado en el modelo supuesto original, \hat{f}_n , es decir, surge un sesgo. En general un λ grande conduce a un sesgo grande, pero debido a tamaños pequeños de $\hat{\mathcal{F}}_\lambda$ conduce a una varianza menor de $\hat{f}_{n,\lambda}$.

Los parámetros, que deben seleccionarse antes del cálculo de \hat{f}_n se denominan generalmente *Hiperparámetros* o *Parámetros de ajuste* sin especificar el propósito. En la observación 7.24 por ejemplo, estos términos también se utilizan para λ .

7.6.4 Entrenamiento, Validación y Prueba

El proceso de calcular una regla de decisión $\hat{f}_n(\omega)$ a partir de un algoritmo dado se llama *entrenamiento*.

Para obtener información y comparaciones de calidad, el riesgo $R(\hat{f}_n)$ de la regla de decisión determinada es usado, que luego actúa simultáneamente como una aproximación del error de generalización $\mathbb{E}R(\hat{f}_n)$. Sin embargo, para especificar $R(\hat{f}_n)$, es necesario el conocimiento de $\mathbb{P}^{(X,Y)}$, lo cual no es el caso en la práctica. Sin embargo, después del cálculo de \hat{f}_n , se ponen a disposición datos de prueba adicionales independientes de los datos de entrenamiento, entonces $R(\hat{f}_n)$ se puede estimar promediando las pérdidas

recibidas:

Nota 7.26. (*Estimación de $R(\hat{f}_n)$*) Sean $(\tilde{X}_i, \tilde{Y}_i) \stackrel{iid}{\sim} \mathbb{P}^{(X,Y)}$, $i = 1, \dots, m$ observaciones (datos de prueba) independientes de T_n . Entonces, debido a la Ley Fuerte de los Grandes Números:

$$\text{empRT}(\hat{f}_n) := \frac{1}{m} \sum_{i=1}^m L(\tilde{Y}, \hat{f}_n(\tilde{X})) \xrightarrow{m \rightarrow \infty} \mathbb{E}[L(Y, \hat{f}_n(X))|T_n] = R(\hat{f}_n) \text{ c.s}$$

empRT(\hat{f}_n) significa Riesgo empírico de \hat{f}_n en los datos de prueba o errores de prueba.

Los datos de prueba no siempre están disponibles en problemas de aplicaciones. La observación 7.26 se puede utilizar convirtiendo las observaciones proporcionadas en datos de entrenamiento $(X_i, Y_i)_{i=1, \dots, n}$ y los datos de prueba $(\tilde{X}_i, \tilde{Y}_i)_{i=1, \dots, m}$ se dividen (por ejemplo, en la relación n:m = 90:10) y el cálculo de \hat{f}_n solo se realiza utilizando los datos de entrenamiento.

Nota 7.27. El riesgo empírico de \hat{f}_n sobre los datos de entrenamiento,

$$\hat{R}_n(\hat{f}_n) = \frac{1}{n} \sum_{i=1}^n L(Y_i, \hat{f}_n(X_i))$$

también se denomina error de entrenamiento. $\hat{R}_n(\hat{f}_n)$ no se puede utilizar para estimar $R(\hat{f}_n)$. Al determinar \hat{f}_n se incorporó información de los datos de entrenamiento; por ejemplo, si se utilizó la observación 7.23, la regla de decisión $\hat{f}_n(\omega)$ se entrenó para adaptarse particularmente bien a $(X_i, Y_i)_{i=1, \dots, n}$. Esto viola el supuesto de independencia de la observación 7.26, y en general, $\hat{R}_n(\hat{f}_n) \leq R(\hat{f}_n)$. En el caso extremo de sobreajuste, Ec.(7.5) $\hat{R}_n(\hat{f}_n) = 0$ (aunque la calidad de \hat{f}_n suele ser muy mala, especialmente en este caso).

Si $\hat{f}_{n,\lambda}$ depende de un parámetro de ajuste $\lambda \geq 0$, entonces, en la práctica, se necesita una guía para una elección bien fundamentada de λ . Nos concentraremos aquí solo en los parámetros de ajuste del tipo Ec.(7.11), pero los enfoques que se presentan a continuación también se pueden utilizar para todos los demás tipos.

Un enfoque común trabaja con la estimación del riesgo $R(\hat{f}_{n,\lambda})$ y una elección de λ basada en esto. Como ya se motivó en la observación 7.26, dividimos las observaciones dadas en el conjunto $(X_i, Y_i)_{i=1,\dots,n}$ (datos de entrenamiento) y un conjunto $(\tilde{X}_i, \tilde{Y}_i)_{i=1,\dots,m}$ (datos de validación) y procedemos de la siguiente manera:

Nota 7.28. (*Determinación de hiperparámetros adecuados: validación*) Sean $(\tilde{X}_i, \tilde{Y}_i) \stackrel{iid}{\sim} \mathbb{P}^{(X,Y)}$, $i = 1, \dots, m$ observaciones (datos de validación) independientes de T_n . Determine $\hat{f}_{n,\lambda}$ basándose en los datos de entrenamiento $(X_i, Y_i)_{i=1,\dots,n}$ y elija

$$\hat{\lambda}^{std} := \arg \min_{\lambda \geq 0} empVR(\hat{f}_{n,\lambda}) \quad (7.12)$$

dado que

$$empVR(\hat{f}_{n,\lambda}) := \frac{1}{m} \sum_{i=1}^m L(\bar{Y}_i, \hat{f}_{n,\lambda}(\bar{X}_i))$$

el riesgo empírico se indica en los datos de validación.

Observaciones:

- ✓ El procedimiento de la observación 7.28 se llama *Validación*. El nombre se refiere al hecho de que los diferentes valores de λ se pueden verificar (validar) sobre los datos de validación y luego se selecciona el valor más adecuado.
- ✓ Esperamos que $\hat{f}_{n,\hat{\lambda}^{std}}$ según la observación 7.26 y
$$R(\hat{f}_{n,\hat{\lambda}^{std}}) \approx empVR(\hat{f}_{n,\hat{\lambda}^{std}}) = \min_{\lambda \geq 0} empVR(\hat{f}_{n,\lambda}) \approx \min_{\lambda \geq 0} R(\hat{f}_{n,\lambda})$$
alcanza al menos aproximadamente el riesgo mínimo posible entre todos los $\hat{f}_{n,\lambda}$.
- ✓ Para llevar a cabo el enfoque anterior, $\hat{f}_{n,\lambda}$ no debe calcularse para todo $\lambda \geq 0$. A menudo, el cálculo es suficiente, por ejemplo, en una cuadrícula geométrica $\lambda \in \{a^k : k \in \mathbb{Z}\}$ con $0 < a < 1$, por lo que deben excluirse valores particularmente grandes y pequeños de λ .

Nota 7.29. Si se determinó mediante validación que $\hat{f}_n = \hat{f}_{n,\hat{\lambda}^{std}}$, ni el riesgo empírico a través de los datos de entrenamiento $\hat{R}_n(\hat{f}_n)$ ni el riesgo empírico a través de los datos de validación $empVR(\hat{f}_n)$ pueden usarse para

estimar $R(\hat{f}_n)$. Como regla general, se aplica lo siguiente: tan pronto como los datos se hayan incorporado al cálculo de \hat{f}_n de alguna forma, ya no se podrán utilizar para determinar una estimación del riesgo $R(\hat{f}_n)$.

Una combinación correcta de entrenamiento, validación y cálculo de riesgos resulta del siguiente procedimiento:

Nota 7.30. (*entrenamiento, validación y prueba*)

Divida las observaciones originales en datos de entrenamiento $(X_i, Y_i)_{i=1,\dots,n}$, datos de validación $(\tilde{X}_i, \tilde{Y}_i)_{i=1,\dots,m_1}$ (si es necesario) y datos de prueba $(\tilde{\tilde{X}}_i, \tilde{\tilde{Y}}_i)_{i=1,\dots,m_2}$, por ejemplo en la relación $n : m_1 : m_2 = 70 : 20 : 10$. Luego, haga lo siguiente:

1. Encuentre $\hat{f}_{n,\lambda}$ con $(X_i, Y_i)_{i=1,\dots,n}$ para diferentes valores de $\lambda \geq 0$.
2. Encuentre $\hat{f}_n := \hat{f}_{n,\hat{\lambda}^{std}}$ de acuerdo con la ecuación (7.12).
3. Una estimación de $R(\hat{f}_n)$ viene dada por el error de prueba (ver observación 1.17)

$$\text{empRT}(\hat{f}_n) = \frac{1}{m_2} \sum_{i=1}^{m_2} L(\tilde{\tilde{Y}}_i, \hat{f}_n(\tilde{\tilde{X}}_i)).$$

El error de entrenamiento viene dado por $\hat{R}_n(\hat{f}_n) = \frac{1}{n} \sum_{i=1}^n L(Y_i, \hat{f}_n(X_i))$ (ver Observación 7.27).

En este libro, al determinar los algoritmos, siempre nos enfocamos solo en las técnicas para el paso (1).

Seguimiento de errores de validación en procesos iterativos

Algunos algoritmos $\hat{f}_{n,\lambda}$ como las redes neuronales o el refuerzo de árboles(boosting) se determinan mediante procesos de iteración, que en cada paso $m \in \mathbb{N}$ entrega un algoritmo $\hat{f}_{n,\lambda}^{(m)}$. En los ejemplos mencionados, el valor límite $\lim_{m \rightarrow \infty} \hat{f}_{n,\lambda}^{(m)} = \hat{f}_{n,\lambda}$ es el algoritmo definido por el proceso de iteración *sobreadaptado* a los datos de entrenamiento, con la sobre adaptación siendo "acumulada" por el proceso de iteración. Esto se hace activando más parámetros de la aceptación del modelo con cada nuevo paso de iteración. El algoritmo $\hat{f}_{n,\lambda}^{(m)}$ satisface entonces $\hat{f}_{n,\lambda}^{(m)} \in \hat{\mathcal{F}}_\lambda^{(m)} \subset \mathcal{F}$, y para $m' > m$ tenemos $\hat{\mathcal{F}}_\lambda^{(m)} \subset \hat{\mathcal{F}}_\lambda^{(m')}$. En consecuencia, uno espera la curva de riesgos

$$m \mapsto R\left(\hat{f}_{n,\lambda}^{(m)}\right)$$

Para un m pequeño inicialmente disminuye (adaptación inicial a los datos de entrenamiento, sesgo decreciente), alcanza un mínimo en un cierto paso de iteración $m^* \in \mathbb{N}$ y luego aumenta nuevamente (debido a la variación creciente), ver Figura 7.3

Al monitorear el riesgo empírico en los datos de validación, se puede determinar una condición de terminación adecuada para el proceso de iteración. Este procedimiento se denomina detención anticipada:

Nota 7.31. (*Seguimiento del error de validación en el caso de procedimientos iterativos*)

Divida las observaciones originales en datos de entrenamiento $(X_i, Y_i)_{i=1, \dots, n}$ y datos de validación $(\bar{X}_i, \bar{Y}_i)_{i=1, \dots, m_1}$. Sea $M \in \mathbb{N}$ un número máximo posible de iteraciones realizadas.

1. Para diferentes $\lambda \geq 0$

(1a) Para cada iteración $m \in \{1, 2, \dots, M\}$ determine el algoritmo $\hat{f}_{n,\lambda}^{(m)}$ y el riesgo empírico asociado

$$\text{empRV}(\hat{f}_{n,\lambda}^{(m)}) = \frac{1}{n} \sum_{i=1}^n L\left(\bar{Y}_i, \hat{f}_{n,\lambda}^{(m)}(\bar{X}_i)\right)$$

(1b) Determine $m^*(\lambda) := \arg \min_{m \in \{1, \dots, M\}} \text{empRV}(\hat{f}_{n,\lambda}^{(m)})$

2. Encuentre $\hat{f}_n := \hat{f}_{n,\hat{\lambda}^{std}}^{m^*(\hat{\lambda}^{std})}$ de acuerdo con la ecuación (7.12) de $\hat{f}_{n,\lambda}^{m^*(\lambda)}$.

Observación: Para la aplicación del procedimiento anterior \hat{f}_n no tiene que depender necesariamente de λ ; el paso 2 luego cambia a $\hat{f}_n := \hat{f}_n^{(m^*)}$

7.6.5 Métodos de validación alternativos

Si $\hat{f}_{n,\lambda}$ depende de un hiperparámetro λ , las observaciones originales deben dividirse en datos de entrenamiento y validación para determinar un algoritmo final $\hat{f}_n = \hat{f}_{n,\hat{\lambda}}$ de acuerdo con la observación 7.30. Solo los n datos de entrenamiento se incluyen en el cálculo de $\hat{f}_{n,\lambda}$; los datos de validación solo se utilizan para seleccionar un λ adecuado.

Para un error de generalización bajo de \hat{f}_n o un riesgo bajo de la función de decisión asociada, el número de datos de entrenamiento n utilizados es decisivo y debe evitarse el uso de demasiadas observaciones para la validación. Por otro lado, no se deben usar muy pocos datos para la validación, porque entonces la calidad de la selección de λ se ve afectada. Este dilema se produce sobre todo cuando hay pocas observaciones para determinar \hat{f}_n .

Aquí presentamos dos métodos que permiten que todas las observaciones se utilicen para determinar $\hat{f}_{n,\lambda}$ a pesar de un hiperparámetro. A diferencia de la Observación 7.30, la selección de λ no se realiza sobre la base del riesgo $R(\hat{f}_{n,\lambda})$ sino sobre la base del error de generalización $\mathbb{E}R(\hat{f}_{n,\lambda})$ (o una estimación del mismo). Este enfoque debe usarse con precaución. En contraste con el riesgo empírico, el error de generalización no tiene en cuenta los datos de entrenamiento disponibles actualmente T_n , pero hace la elección sobre la base del riesgo promedio sobre todas las posibles configuraciones observables de datos de entrenamiento. Esto significa que el $\hat{\lambda}$ seleccionado es una buena opción para muchas configuraciones de datos de entrenamiento, pero puede que no lo sea para el actual. Discutimos brevemente la influencia de este hecho en la calidad de los procedimientos, así como otras ventajas y desventajas en las observaciones 7.33 y 7.39.

Validación cruzada (Cross Validation)

En el caso de la validación cruzada, ningún dato de validación $(\bar{X}_i, \bar{Y}_i)_{i=1,\dots,m_1}$ se separa de las observaciones. En cambio, los datos de entrenamiento se dividen en M grupos de igual tamaño (Ver Figura 7.5) y se utilizan para validación entre ellos.

Para llevar a cabo el método, el algoritmo no solo debe ser calculable para exactamente n datos de entrenamiento, sino que se requiere una regla de educación general $\hat{f}_{N,\lambda} = A_{N,\lambda}(T_N)$ del algoritmo para cualquier número $N \in \mathbb{N}$ de datos de entrenamiento T_N . En particular, debe asegurarse que λ tenga el mismo significado y la misma influencia en la apariencia del algoritmo en esta regla de formación, independientemente del número de datos de entrenamiento. La regla de educación Ec.(7.11), por ejemplo, permite directamente la estimación de $\hat{f}_{N,\lambda}$ para cualquier número de datos de entrenamiento, y λ siempre representa el castigo de una propiedad indeseable (en la misma cantidad).

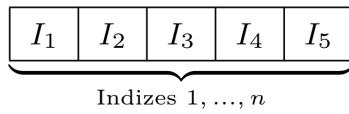


Figura 7.5: Distribución de datos de entrenamiento con validación cruzada de 5 folds(pliegues)

Para reducir la carga de notación, suprimimos la dependencia de A de N en la siguiente definición, es decir, escribimos $A_\lambda(T_N) = A_{N,\lambda}(T_N)$.

Definición 7.32. (*Validación cruzada M-Fold*)

Sea $M \in \{2, \dots, n\}$.

(0) *Divida los datos de entrenamiento al azar en M partes iguales. Obtenemos una descomposición disyunta.*

$$\{1, \dots, n\} = \bigcup_{m=1}^M I_m, \quad I_m \subset \{1, \dots, n\} \quad (m = 1, \dots, M)$$

(1') *Calcule el algoritmo $\tilde{f}_{n,\lambda^{(-m)}} := A_\lambda(T_n^{(-m)})$ basado en los datos de entrenamiento $T_n^{(-m)} := (X_i, Y_i)_{i \in \{1, \dots, n\} / I_m}$ para diferentes $\lambda \geq 0$.*

(2') *Sea $\hat{f}_n = \hat{f}_{n,\hat{\lambda}^{cv}}(T_n)$ el algoritmo basado en todos los datos de entrenamiento, donde*

$$\hat{\lambda}^{cv} := \arg_{\lambda \geq 0} \min CV_n(\lambda),$$

y

$$CV_n(\lambda) := \frac{1}{n} \sum_{i=1}^n L \left(Y_i, \tilde{f}_{n,\lambda}^{(-m(i))}(X_i) \right) = \frac{1}{n} \sum_{m=1}^M \sum_{i \in I_m} L \left(Y_i, \tilde{f}_{n,\lambda}^{(-m)}(X_i) \right)$$

y $m : \{1, \dots, n\} \rightarrow \{1, \dots, M\}$ asigna el índice m a cada índice i con $i \in I_m$.

En este caso escribimos que \hat{f}_n o $\hat{\lambda}^{cv}$ se eligieron mediante una validación cruzada M-Fold (engl. M-fold Cross Validation).

Nota 7.33. (*Sobre la teoría de la validación cruzada*)

- ✓ En el caso que $M = n$ el entrenamiento corresponde a n algoritmos en los que se omite una observación en cada caso. Este caso especial se llama validación cruzada dejando uno fuera.
- ✓ Con la validación cruzada M -Fold, el riesgo empírico de los algoritmos $\tilde{f}_{n,\lambda}^{(-m)}$, $m=1,\dots,M$ basados en $T_n^{(.m)}$ se determina con los datos de entrenamiento omitidos. $CV_n(\lambda)$ promedia estas expresiones para $m=1,\dots,M$. Por lo tanto, a diferencia de $empVR(\hat{f}_{n,\lambda})$ en la observación 1.19, $CV_n(\lambda)$ no representa una estimación directa de $R(\hat{f}_{n,\lambda})$. Suponiendo por simplicidad que $\#I_m = \frac{n}{M} \in \mathbb{N}$, esperamos, debido a la Ley de los Grandes Números, que:

$$\begin{aligned} CV_n(\lambda) &= \frac{1}{M} \sum_{i=1}^M \frac{1}{\#I_m} \sum_{i \in I_m} L \left(Y_i, \tilde{f}_{n,\lambda}^{(-m)}(X_i) \right) \\ &\approx \frac{1}{M} \sum_{i=1}^M R \left(\tilde{f}_{n,\lambda}^{(-m)} \right) \approx \mathbb{E}R \left(\tilde{f}_{n,\lambda}^{(-1)} \right) = \mathbb{E}R(\hat{f}_{n-\frac{n}{M},\lambda}) \end{aligned} \quad (7.13)$$

Es decir, $CV_n(\lambda)$ estima el error de generalización de $\hat{f}_{n-\frac{n}{M},\lambda}$ en lugar del riesgo $R(\hat{f}_{n,\lambda})$. Así, en los casos extremos $M \in \{2, n\}$, se elige realmente un λ óptimo para un algoritmo basado en $\frac{n}{2}$ o $n-1$ datos de entrenamiento; por tanto, desde un punto de vista teórico, es preferible un M alto.

- ✓ El hecho de que $CV_n(\lambda)$ estime un error de generalización no es un problema aquí, ya que la estimación se realiza con los datos de entrenamiento sin ninguna otra suposición del modelo. En este sentido, cabe esperar que la minimización de $CV_n(\lambda)$ produzca resultados tan buenos que tengan en cuenta los datos de entrenamiento como la minimización de $R(\hat{f}_{n,\lambda})$ en la observación 7.28.

Nota 7.34. (Sobre la aplicación de la validación cruzada) En la práctica, a menudo se elige M pequeño, por ejemplo, $M = 5$ o $M = 10$. La razón es que para el rendimiento de validación cruzada, el algoritmo debe ser determinado M veces con diferentes datos de entrenamiento y para diferentes λ . Si la determinación de los algoritmos considerados requiere mucho tiempo, la validación cruzada a menudo no puede aplicarse en la práctica.

AIC - Criterio de Información de Akaike

El Criterio de información de Akaike (AIC) fue desarrollado por [1]. Estima el error de generalización $\mathbb{E}R(\hat{f}_{n,\lambda})$ corrigiendo el error de entrenamiento

$\hat{R}_n(\hat{f}_{n,\lambda})$. Aquí discutimos el método solo de forma heurística y para un problema específico de regresión, ya que rara vez se utiliza en la práctica del aprendizaje automático:

Nota 7.35. (AIC) Sean $X_1, \dots, X_n \in \mathcal{X} \subset \mathbb{R}^d$ puntos de medición predeterminados y no aleatorios. Sea $f^* : \mathcal{X} \rightarrow \mathbb{R}$ una función medible. Para los datos $(X_i, Y_i)_{i=1, \dots, n}$ se mantienen:

$$\varepsilon_i := Y_i - f^*(X_i), \quad i = 1, \dots, n \text{ son variables aleatorias i.i.d reales con } \mathbb{E}\varepsilon_i = 0. \quad (7.14)$$

Nota 7.36. Dado que los X_i son deterministas, $(X_i, Y_i)_{i=1, \dots, n}$ no son i.i.d, por lo que los resultados de convergencia para los errores de entrenamiento y generalización no son directamente transferibles. Sin embargo, por los supuestos del modelo $f^*(x) = \mathbb{E}[Y|X = x]$.

En este contexto, el error de entrenamiento de cualquier $\hat{f}_{n,\lambda}$ es:

$$\hat{R}_n(\hat{f}_{n,\lambda}) = \frac{1}{n} \sum_{i=1}^n L\left(Y_i, \hat{f}_{n,\lambda}(X_i)\right).$$

Sean (X_i, \tilde{Y}_i) , $i=1, \dots, n$ nuevas observaciones independientes de (X_i, Y_i) , $i=1, \dots, n$ que siguen la Ec.(7.14). El error de generalización de $\hat{f}_{n,\lambda}$ es entonces $\mathbb{E}R(\hat{f}_{n,\lambda})$ con:

$$R(\hat{f}_{n,\lambda}) = \frac{1}{n} \sum_{i=1}^n L\left(\tilde{Y}_i, \hat{f}_{n,\lambda}(X_i)\right).$$

Existe la siguiente relación entre estas dos variables:

Lema 7.37. (Error de entrenamiento vs. Error de generalización) Sea $L(y, s) = (y - s)^2$ la función de pérdida cuadrática. Entonces se aplica lo siguiente:

$$\mathbb{E}R(\hat{f}_{n,\lambda}) - \mathbb{E}\hat{R}_n(\hat{f}_{n,\lambda}) = \frac{2}{n} \sum_{i=1}^n \text{Cov}\left(Y_i, \hat{f}_{n,\lambda}(X_i)\right)$$

Prueba: Se da

$$\begin{aligned} &= [\tilde{Y}_i^2 - 2\tilde{Y}_i \hat{f}_{n,\lambda}(X_i) + \hat{f}_{n,\lambda}(X_i)^2] - [Y_i^2 - 2Y_i \hat{f}_{n,\lambda}(X_i) + \hat{f}_{n,\lambda}(X_i)^2] \\ &= \tilde{Y}_i^2 - Y_i^2 + 2(Y_i \hat{f}_{n,\lambda}(X_i) - \tilde{Y}_i \hat{f}_{n,\lambda}(X_i)). \end{aligned} \quad (7.15)$$

Entonces sigue

$$\begin{aligned} \mathbb{E}R(\hat{f}_{n,\lambda}) - \mathbb{E}\hat{R}_n(\hat{f}_{n,\lambda}) &= \frac{1}{n} \sum_{i=1}^n \left\{ \mathbb{E} \left(\tilde{Y}_i - \hat{f}_{n,\lambda}(X_i) \right)^2 - \mathbb{E} \left(Y_i - \hat{f}_{n,\lambda}(X_i) \right)^2 \right\} \\ &\stackrel{(2.11)}{=} \frac{2}{n} \sum_{i=1}^n \left\{ \mathbb{E} \left[Y_i \hat{f}_{n,\lambda}(X_i) \right] - \underbrace{\mathbb{E} \left[\tilde{Y}_i \hat{f}_{n,\lambda}(X_i) \right]}_{=\mathbb{E}\tilde{Y}_i \mathbb{E}\hat{f}_{n,\lambda}(X_i)} \right\} \quad (7.17) \\ &\stackrel{\mathbb{E}Y_i = \mathbb{E}\tilde{Y}_i}{=} \frac{2}{n} \sum_{i=1}^n Cov(Y_i, \hat{f}_{n,\lambda}(X_i)). \end{aligned}$$

Si utilizamos la aproximación $\mathbb{E}\hat{R}_n(\hat{f}_{n,\lambda}) \approx \hat{R}_n(\hat{f}_{n,\lambda})$, obtenemos del Lema 1.26:

$$\mathbb{E}R(\hat{f}_{n,\lambda}) \approx \hat{R}_n(\hat{f}_{n,\lambda}) + \frac{2}{n} \sum_{i=1}^n Cov(Y_i, \hat{f}_{n,\lambda}(X_i)) \quad (7.18)$$

Esto significa que se puede obtener una estimación del error de generalización corrigiendo el error de entrenamiento con un sumando adicional. Los términos $Cov(Y_i, \hat{f}_{n,\lambda}(X_i))$ representan la dependencia entre Y_i y la predicción $\hat{f}_{n,\lambda}(X_i)$ y se hacen mayores para $Y_i = \hat{f}_{n,\lambda}(X_i)$. En este sentido, el sumando adicional penaliza el sobre ajuste a los datos de entrenamiento.

La Ec.(7.18) carece de valor en la práctica sin un supuesto de modelo más detallado, ya que no existen estimadores sencillos para la cantidad teórica $Cov(Y_i, \hat{f}_{n,\lambda}(X_i))$. Sin embargo, para supuestos de modelos simplemente estructurados, $\sum_{i=1}^n Cov(Y_i, \hat{f}_{n,\lambda}(X_i))$ puede calcularse y estimarse explícitamente. Una forma frecuente es

$$Cov(Y_i, \hat{f}_{n,\lambda}(X_i)) = df(\lambda) \cdot \sigma_\varepsilon^2, \quad (7.19)$$

con $df(\lambda) \geq 0$ que solo depende de λ y $\sigma_\varepsilon^2 := Var(\varepsilon_1)$. Entonces $df(\lambda)$ se llama el *número de grados de libertad* o el *número efectivo de parámetros* en el modelo. En este caso, la validación de λ puede llevarse a cabo de la siguiente manera:

Definición 7.38. (*Validación del AIC*) Suponga que se cumplen las hipótesis del modelo 7.35 y la ecuación (7.19). Sea $\hat{\sigma}_n^2$. Entonces se llama:

$$AIC_n(\lambda) := \hat{R}_n \left(\hat{f}_{n,\lambda} \right) + \frac{2}{n} \cdot df(\lambda) \cdot \sigma_\varepsilon^2 \quad (7.20)$$

Criterio de Información de Akaike y $\hat{\lambda}^{aic} := \arg \min_{\lambda \geq 0} AIC_n(\lambda)$. En este caso escribimos que $\hat{\lambda}^{aic}$ fue elegido por la validación del AIC.

Nota 7.39. ✓ A diferencia de la validación cruzada para determinar $AIC(\lambda)$ es necesario un supuesto de modelo concreto debido a la ecuación (7.19). Si esto es así, el procedimiento arroja un buen $\hat{\lambda}^{aic}$ que, sin embargo, puede no ajustarse de forma óptima a la función de decisión $\hat{f}_{n,\lambda}(\omega)$ que se acaba de determinar a partir de los datos de entrenamiento. Véanse las observaciones al principio de Sección 1.5.

- ✓ Una aproximación general a la elección de $\hat{\sigma}_n^2$ es $\hat{\sigma}_n^2 = \hat{R}_n(\hat{f}_{n,\tilde{\lambda}})$, donde $\lambda \geq 0$ debe elegirse de forma que al menos no se haya producido todavía ningún sobre ajuste debido a $\hat{f}_{n,\tilde{\lambda}}$. Por lo tanto, al utilizar este estimador, hay que conocer ya un rango aproximado en el que λ debe estar.
- ✓ La calidad de $\hat{\lambda}^{aic}$ depende de la calidad del estimador de $\hat{\sigma}_n^2$ y en la calidad de los supuestos del modelo (con respecto al error de aproximación) o la ecuación resultante (7.19).

Utilizaremos el AIC solo una vez en el capítulo 2 en el estimador de cresta. Las modificaciones del AIC, desarrolladas para los modelos bayesianos en el aprendizaje automático, son, por ejemplo, el BIC (*Criterio de Información de Bayes*) o el DIC (*Criterio de Información de Desviación*).

7.6.6 Procesamiento de datos en la práctica

En los problemas de aplicación, los componentes individuales de los datos X_i pueden diferir en varios órdenes de magnitud. Aunque esto sea irrelevante para la teoría matemática, esto puede ser problemático a la hora de calcular los algoritmos en el ordenador. En algunos algoritmos, por ejemplo, se aplican a los datos de entrenamiento X_i operaciones aritméticas como funciones exponenciales o funciones polinómicas de alto grado, de modo que valores demasiado grandes o demasiado pequeños de los X_i pueden provocar errores en la aritmética de coma flotante o desbordamientos aritméticos de las variables utilizadas.

Por ello, es aconsejable normalizar los datos antes de aplicar los algoritmos y posiblemente ampliarlos para incluir variables deterministas. Formalmente lo expresamos mediante una transformación

$$T : \mathcal{X}^n \rightarrow \tilde{\mathcal{X}}^n,$$

que mapea los datos de entrenamiento originales $(X_{ij})_{j=1,\dots,d} = X_i \in \mathcal{X} \subset \mathbb{R}^d$ a los nuevos datos de entrenamiento $(\tilde{X}_{ij})_{j=1,\dots,d} = \tilde{X}_i = T(X_i) \in \tilde{\mathcal{X}} \subset \mathbb{R}^{\tilde{d}}$. Algunas transformaciones típicas son:

Ejemplo 7.40. (Ejemplos de transformaciones de preprocesamiento)

- (i) Centrado y normalización (genera datos de entrenamiento \tilde{X}_i con magnitudes iguales):

$$X_{ij} := \frac{X_{ij}^o}{\sqrt{\frac{1}{n} \sum_{k=1}^n (X_{kj}^o)^2}}, \quad X_{ij}^o := X_{ij} - \frac{1}{n} \sum_{k=1}^n X_{kj}$$

Después de esta transformación $\frac{1}{n} \sum_{i=1}^n \tilde{X}_{ij} = 0$ y $\frac{1}{n} \sum_{i=1}^n \tilde{X}_{ij}^2 = 1$.

- (ii) Los supuestos demasiado restrictivos de modelos individuales se pueden expandir modificando los datos de entrenamiento X_i e incrustándolos en un espacio más grande, luego aplicando los algoritmos de los modelos a (\tilde{X}_i, Y_i) en lugar de (X_i, Y_i) . Dos ejemplos de T son:

$$\tilde{X}_i := (1, (X_i)^T)^T \quad (\text{Añadiendo una media}) \tag{7.21}$$

$\tilde{X}_i := (1, X_{i1}, \dots, X_{id}, X_{i1}^2, \dots, X_{id}^2)$ (Funciones cuadráticas de los componentes)

Formalmente, una extensión del modelo de este tipo significa que la hipótesis $f^* \in \mathcal{F}$ realizada anteriormente se modifica por $f^* = \tilde{f}(T(x))$ con $\tilde{f}^* \in \mathcal{F}$. Un algoritmo obtenido mediante $(\tilde{X}_i, Y_i) \tilde{f}_n : \tilde{\mathcal{X}} \rightarrow \mathcal{Y}$ por lo tanto, produce un algoritmo para las observaciones originales $x \in \mathcal{X}$ usando $\hat{f}_n(x) := \tilde{f}_n(T(x))$.

Atención: En particular con respecto a (ii) se aplica lo siguiente:

- Al aplicar un algoritmo a los datos transformados $(\tilde{X}_i, Y_i)_{i=1,\dots,n}$ el supuesto del modelo debe aplicarse a estos y ya no a los datos originales $(X_i, Y_i)_{i=1,\dots,n}$!
- Dado que muchos algoritmos funcionan peor con dimensiones altas de \tilde{X}_i , este La técnica debe utilizarse con precaución. En el capítulo 4 conoceremos una extensión.

Hay un resultado teórico que permite la manipulación de observaciones con autocorrelación temporal con ciertas restricciones, ver [22]

Nota 7.41. No-free-Lunch. *Siguiendo a [13] existe una familia de teorías que en donde se plantean que no hay algoritmos o procesos de optimización, etc. tal que sean superiores a otros en todos los problemas o casos. Así lo explican*

" El teorema *No free-lunch para el aprendizaje automático* [42] establece que, promediando todas las posibles distribuciones de generación de datos, cada algoritmo por ejemplo de clasificación tiene la misma tasa de error al clasificar puntos previamente no observados. En otras palabras, en cierto sentido, ningún algoritmo de aprendizaje automático es universalmente mejor que otro. El algoritmo más sofisticado que podemos concebir tiene el mismo rendimiento promedio (sobre todas las tareas posibles) que simplemente predecir que cada punto pertenece a la misma clase.

Afortunadamente, estos resultados solo se mantienen cuando promediemos todas las posibles distribuciones de generación de datos. Si hacemos suposiciones sobre los tipos de distribuciones de probabilidad que encontramos en las aplicaciones del mundo real, podemos diseñar algoritmos de aprendizaje que funcionen bien en estas distribuciones.

Esto significa que el objetivo de la investigación del aprendizaje automático no es buscar un algoritmo de aprendizaje universal o el mejor algoritmo de aprendizaje absoluto. En cambio, nuestro objetivo es comprender qué tipos de distribuciones son relevantes para el "mundo real" que experimenta un agente de IA y qué tipos de algoritmos de aprendizaje automático funcionan bien con los datos extraídos de los tipos de distribuciones de generación de datos que nos interesan. "

7.7 Árboles, Redes Neuronales y Aprendizaje Profundo

En ésta sección introduciremos las ideas de redes neuronales para el análisis de series de tiempos, más específicamente, para hacer predicción. Primero hablaremos acerca de los Árboles , luego de los perceptrones multicapa o la red neuronal multicapa, y después introduciremos las redes Long-Short-Term Multicapa(LSTM) como un caso particular de las redes neuronales recurrentes(RNN) para el análisis de series de tiempo. Note que estos son enfoques son procedimientos estadísticos semi-paramétricos ver [38].

7.7.1 Métodos basados en Árboles

Vamos a seguir los lineamientos del libro [38]. Los métodos basados en árboles han sido bastante utilizados para predicción y clasificación, especialmente para observaciones independientes. Hay bastantes métodos basados en árboles, árboles de regresión, árboles de clasificación, sin embargo la idea fundamental, es la misma, la cual consiste en la estratificación o segmentación del espacio de predictores dentro del múltiples subregiones, donde cada de las cuales contiene relativamente observaciones mas homogéneas, tal que modelos simples pueden ser usados en esas subregiones.

La idea consiste en configurar el problema resolver como un problema de aprendizaje supervisado. Por ejemplo supongamos que se tiene una serie de tiempo de tamaño $T = 7$, es decir $\{y_1, y_2, y_3, y_4, y_5, y_6, y_7\}$, y que se desea por ejemplo predecir la siguiente observación y_{t+1} basado en las $k = 2$ pasadas observaciones y_t, y_{t-1} . Entonces la configuración del problema de aprendizaje supervisado(**para la predicción 1-paso adelante basados en las dos pasadas observaciones**) quedaría de la siguiente forma:

y_{t+1}	y_t, y_{t-1}	\underline{x}_{t+1}
y_3	y_2, y_1	\underline{x}_3
y_4	y_3, y_2	\underline{x}_4
\vdots	\vdots	
y_7	y_6, y_5	\underline{x}_7

Supongamos ahora que se desea predecir la observación y_{t+2} basado en las $k = 3$ pasadas observaciones y_t, y_{t-1}, y_{t-2} . Entonces la configuración del problema de aprendizaje supervisado(**para la predicción 2-pasos adelante basados en las tres pasados observaciones**) quedaría de la siguiente forma:

y_{t+2}	y_t, y_{t-1}, y_{t-2}	\underline{x}_{t+2}
y_5	y_4, y_3, y_2	\underline{x}_4
y_6	y_4, y_3, y_2	\underline{x}_5
y_7	y_5, y_4, y_3	\underline{x}_6

7.7.2 Árboles de decisión

Los árboles binarios son el componente básico de la mayoría de los métodos estadísticos basados en árboles. Binario quiere decir, que cada rama del árbol puede ser dividido dentro de 2 sub-ramas Estos árboles son comúnmente

referidos a árboles de decisión. Veamos un ejemplo ilustrativo; consideremos las tasas de crecimiento trimestral en porcentajes, del producto interno bruto de los Estados Unidos desde 1947.II hasta 2015.II, para un total de 273 observaciones. Vea la gráfica 7.6 para la serie de crecimientos del PIB de EE.UU.

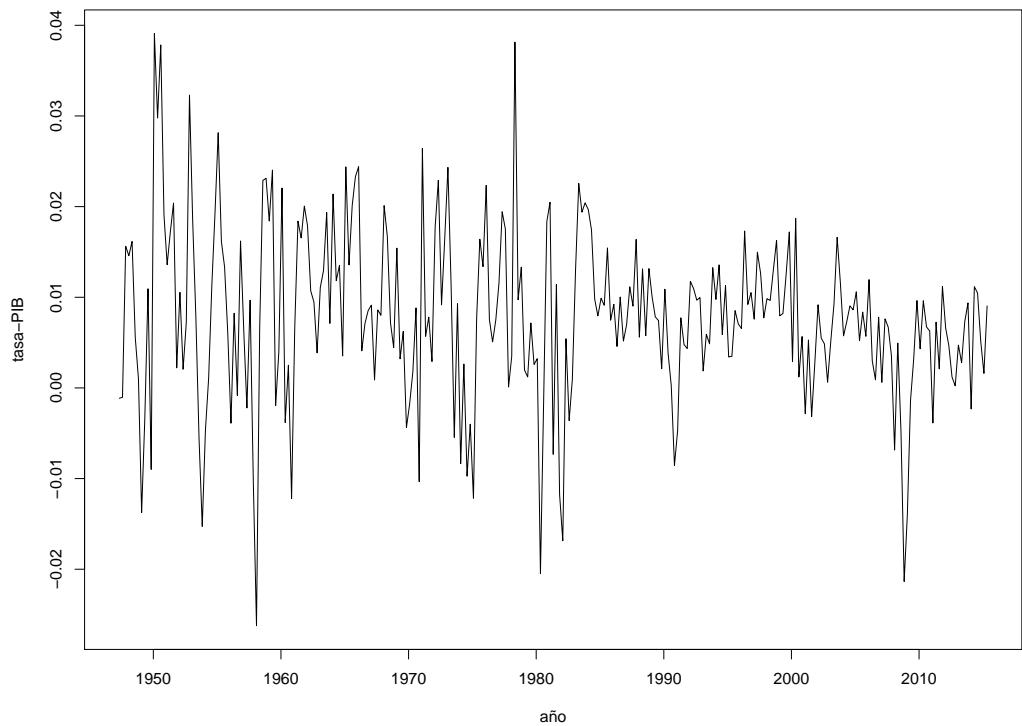


Figura 7.6: Tasa de Crecimiento del PIB US

La idea consistirá en predecir la tasa de crecimiento del PIB trimestral, usando los valores pasados de las misma tasas de crecimiento, es decir, el modelo subyacente propuesto será

$$y_t = g(y_{t-1}, y_{t-2}, y_{t-3}) + a_t$$

donde y_t es la tasa de crecimiento del PIB en el tiempo t , a_t denota el término de error, y $g(\cdot)$ es una función suave de \mathbb{R}^3 a \mathbb{R} . El árbol de decisión representa una aproximación no-lineal a trozos de $g(\cdot)$.

Basados en ese problema de aprendizaje supervisado, lo que tenemos es la siguiente configuración de datos:

y_{t+1}	y_t, y_{t-1}, y_{t-2}	\underline{x}_{t+1}
y_4	y_3, y_2, y_1	\underline{x}_4
y_5	y_4, y_3, y_2	\underline{x}_5
y_6	y_5, y_4, y_3	\underline{x}_6
\vdots	\vdots	\vdots
y_T	$y_{T-1}, y_{T-2}, y_{T-3}$	\underline{x}_T

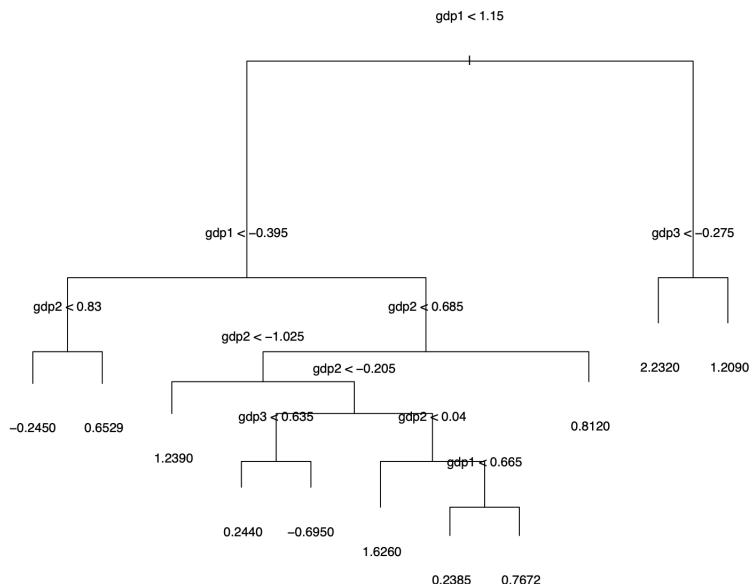


Figura 7.7: Árbol de Decisión para las tasas de crecimiento del PIB US

Note que en la gráfica 7.7 se encuentra el árbol que se encontró. La gráfica fue dibujado boca abajo, de tal manera que las hojas están en la parte inferior del árbol. Para éste ejemplo, el árbol tiene 11 hojas, las cuales muestran un valor numérico(de izquierda a derecha : $-0, 2450, 0.6529, \dots, 2.2320, 1.2090$). Empezando desde arriba, podemos ver que la primera partición es dada por la variable $y_{t-1}(gdp1)$. La rama izquierda es para $y_{t-1} < 1.15$, mientras que la rama derecha es dada por $y_{t-1} \geq 1.5$, que a su vez tiene una subdivisión determinada por la variable $y_{t-3}(gdp3)$.

De la misma manera, podemos ver que la rama izquierda del árbol es subdividida en dos sub-ramas por la misma variable y_{t-1} , cuya sub-sub-rama izquierda es determinada por $y_{t-1} < -0.395$. Así mismo, ésta sub-sub-rama izquierda es subdividida en dos nuevas sub-sub-sub-ramas por la variable y_{t-2} con criterio $y_{t-2} < 0.83$, resultando en dos hojas con valores -0.2450 y 0.6529 . La sub-sub-rama derecha de la rama izquierda tiene una subdivisión determinada también por la variable y_{t-1} , con criterio $y_{t-2} \geq 0.685$, el cual produce una hoja con valor 0.8120 , que es el promedio de las observaciones que satisfacen la condición dada. El resto del árbol puede leerse de manera análoga.

El árbol construido por supuesto puede ser utilizado para predicción. Por ejemplo, supongamos que tenemos $(y_n, y_{n-1}, y_{n-2}) = (0.51, 1.05, 1.12)$ en el origen para pronóstico en $t = n$, y queremos predecir y_{n+1} . Del árbol ajustado, podemos ver que la predicción es 0.812 , puesto que para $t = n + 1$, el valor de la variable rezagada un periodo es $y_n = 0.51$, el cual es menor que 1.15 , así que y_{n+1} pertenece a la rama izquierda de la primera partición. Ahora, como la sub-rama tiene un subdivisión de nuevo basada en y_{t-1} , pero ahora con valor de criterio -0.395 , note que debemos tomar la sub-sub-rama derecha ya que y_n ahora es mayor que -0.395 . Finalmente, tenemos que el segundo rezago y_{n+1} es $y_{n-1} = 1.05$, el cual pertenece a la sub-sub-sub-rama derecha puesto que es mayor que 0.685 . Ese lado derecho ya no tiene mas subdivisiones, por lo tanto termina en una hoja cuyo valor es 0.812 , y así es la predicción para y_{n+1} , la cual corresponde a la media muestral de todos los datos en la misma sub-región que y_{n+1} .

El modelo de árboles puede ser interpretado como sigue: la tasa de crecimiento del PIB de rezago 1 es la variable más importante para determinar el valor de la tasa de crecimiento del PIB. Cuando la tasa de crecimiento del PIB es superior o al igual 1.15% , entonces el retardo 2 de la variable resulta irrelevante (note que no aparece al lado derecho del árbol). Así mismo, la tasa de crecimiento es determinada finalmente por el retardo 3 de la variable. Si la tasa de crecimiento del PIB tres periodos atrás es menor que -0.275 (es decir la economía estuvo en una profunda contracción), entonces se espera que la economía de los Estados Unidos crezca rápido a una tasa del 2.23% . Se puede dar una interpretación análoga a las otras ramas.

Veamos ahora como se construye el árbol continuando con el ejemplo. Recordemos que la primera división se hizo basados en el primer rezago de la variable y_t , es decir y_{t-1} . La división también se basa en un valor de

umbral c , el cual puede ser determinado por el siguiente método: Sea c un número real tal que $y_{t-1(1)} \leq c \leq y_{t-1(n)}$, y sea $R_1 = \{t|y_{t-1} < c\}$ y $R_2 = \{t|y_{t-1} \geq c\}$. Además, sean $\bar{y}_{1,c}$ y $\bar{y}_{2,c}$ las medias muestrales de y_t en R_1 y R_2 respectivamente. La suma de cuadrados del error(SSE) correspondiente al valor de umbral c dado es

$$SSE_1(c) = \sum_{t \in R_1} (y_t - \bar{y}_{1,c})^2 + \sum_{t \in R_2} (y_t - \bar{y}_{2,c})^2, \quad (7.22)$$

la cual se puede ver como una función de c . Note que la primera subdivisión es obtenida por

$$1.15 = \arg \min_c SSE_1(c).$$

Vale la pena decir, que las medias juegan el papel de las predicciones. Note que para el ejemplo, tenemos 3 variables explicativas, y es necesario hacer selección de variables para llevar a cabo las divisiones. Por ejemplo la primera división se hizo basados en y_{t-1} ya que, al hacer la búsqueda por separado con las otras dos variable y_{t-2} y y_{t-3} , y_{t-1} fue elegida porque fue la que minimizó el SS_1 .

Para el siguiente paso vamos a considerar las regiones R_1 y R_2 separadamente. Vamos a considerar la región R_2 , es decir la rama derecha del árbol de la primera división. Lo que considera aquí es que las observaciones en la región R_2 son una nueva muestra y se aplica el mismo procedimiento anterior(tree-growing) para seleccionar la mejor variable explicativa y el mejor valor de umbral para formar una nueva división. Para la región R_2 del ejemplo, ésta es dividida dentro de $R_{21} = \{t|(y_{t-1} \geq 1.15) \cap (y_{t-3} < -0.275)\}$ y $R_{22} = \{t|(y_{t-1} \geq 1.15) \cap (y_{t-3} \geq -0.275)\}$. Este proceso se hace también para la rama izquierda del árbol, es decir, para la región R_1 .

Note que al usar como función objetivo a la suma de cuadrados, el crecimiento del árbol puede continuar si se repite el proceso para seleccionar el mejor predictor y el mejor valor de umbral para la división, de tal manera que se minimice la función objetivo. Sin embargo, un árbol con muchas hojas puede resultar en un sobreajuste, lo cual puede implicar en modelo con bajo poder predictivo. Por lo tanto un proceso de poda es necesario.

7.7.3 Poda del Árbol

Sea \mathcal{T} un árbol con $|\mathcal{T}|$ hojas, el cual es le número de particiones mas uno(o puede depender del número de jerarquías del árbol). Sea R_i la i-ésima región

del árbol para $i = 1, \dots, |\mathcal{T}|$. Además, sea y_t la t -ésima observación de la variable dependiente y \mathbf{x}_t el correspondiente vector de predictores. Sea \bar{y}_{R_i} la media muestral de la variable dependiente de la región R_i

$$\bar{y}_{R_i} = \frac{1}{|R_i|} \sum_t y_t I(\mathbf{x}_t \in R_i)$$

donde $|R_i|$ denota el número de puntos en la región R_i . La idea de la poda del árbol consiste en aplicar una penalización al árbol basado en su complejidad. La complejidad del árbol es medido en términos del número de hojas.

Supongamos que el tamaño de la muestra es T tal que los datos son (y_t, \mathbf{x}_t) para $t = 1, \dots, T$. La selección de modelos usada acá será basada en la predicción fuera de la muestra. Con éste objetivo, se dividen los datos en una submuestra de entrenamiento y una submuestra de pronóstico(prueba). Para el caso de series de tiempo, la submuestra de entrenamiento consiste de los N primeros puntos de datos, mientras que submuestra de pronóstico son los últimos $T - N$ puntos de datos.

Use la muestra de entrenamiento para hacer crecer un gran árbol, digamos \mathcal{T}_0 , el cual se determina con frecuencia estableciendo una cota superior para $|R_i|$, el número de datos en la región R_i . Para un valor de penalidad dada por α , existe un sub-árbol $\mathcal{T} \subset \mathcal{T}_0$ tal que

$$\sum_{t=1}^{|\mathcal{T}|} \sum_{\mathbf{x}_t \in R_i} (y_t - \bar{y}_{R_i})^2 + \alpha |\mathcal{T}|, \quad (7.23)$$

es tan pequeño como sea posible. La ecuación anterior muestra que el parámetro de penalización α controla una compensación entre la complejidad y la bondad del ajuste del subárbol \mathcal{T} . Vamos a denotar éste subárbol con \mathcal{T}_α .

Luego aplicamos el subárbol seleccionado \mathcal{T}_α para producir predicciones en la submuestra de pronóstico o de prueba y calculamos la media de los errores de pronóstico al cuadrado, que también es una función de α . Sea α_0 el valor de α que da la media de los errores de pronóstico al cuadrado mínimos. Finalmente, elegimos \mathcal{T}_{α_0} como el árbol para y_t seleccionado.

7.7.4 Aprendizaje Automático

En éste capítulo nos basaremos del libro [6] y [13]. La inteligencia artificial(IA) y el aprendizaje automático(machine learning) nace en 1955 cuando

un grupo de científicos liderado por el profesor de matemáticas Jhon McCarthy, propusieron un objetivo ambicioso :"El estudio se realizará sobre la base de la conjetura de que todos los aspectos del aprendizaje o cualquier otra característica de la inteligencia pueden, en principio, describirse con tanta precisión que se puede hacer que una máquina los simule. Se intentará encontrar cómo hacer que las máquinas usen el lenguaje, formen abstracciones y conceptos, resuelvan tipos de problemas ahora reservados para los humanos y se mejoren a sí mismos "

Desde entonces, la IA se ha esforzado constantemente por superar a los humanos en varias tareas de evaluación. La métrica más fundamental para este éxito es la prueba de Turing, una prueba de la capacidad de una máquina para exhibir un comportamiento inteligente equivalente o indistinguible de un ser humano. Hay muchos ejemplos de ésto basados en el aprendizaje profundo)(Deep Learning), DeepMind desarrollado por Google, procesamiento de imágenes, neurociencia, etc. En todos estos ejemplos, tenemos que estamos en presencia de grandes cantidades de información y se evaluaron un gran número de modelos candidatos.

Una de las razones por las que la IA y el conjunto de algoritmos informáticos para el aprendizaje, denominados "aprendizaje automático", han tenido éxito es el resultado de una serie de factores que van más allá de los avances en el hardware y el software. Las máquinas pueden modelar procesos de generación de datos complejos y de alta dimensión, barrer millones de configuraciones de modelos y luego evaluar y corregir de manera sólida los modelos en respuesta a nueva información. Note que hay diferentes clases de algoritmos en el área del aprendizaje automático: Aprendizaje Supervisado, Aprendizaje no-supervisado y aprendizaje por refuerzo(reforzado). Los objetivos y configuraciones de los datos difieren para cada clase. En el **aprendizaje supervisado**, se tiene que para un conjunto de datos etiquetados, es decir pares, $(x_1, y_1), (x_2, y_2) \dots, (x_n, y_n)$, cada uno perteneciendo a un espacio de estado particular, el objetivo es aprender la relación que existe entre X y Y . x_i es llamado vector de características y y_i es llamada la etiqueta o respuesta. En el **aprendizaje no-supervisado** un conjunto de datos no etiquetados x_1, x_2, \dots, x_n son dados, y el objetivo consiste en explorar la información para quizás hacer agrupamiento de observaciones similares o encontrar patrones ocultos. Finalmente, el aprendizaje por reforzamiento(Aprendizaje Semi-Supervisado), es un enfoque algorítmico para hacer cumplir la optimización de Bellman de un proceso de decisión de Markov,

que define un conjunto de estados y acciones en respuesta a un régimen cambiante para maximizar alguna noción de recompensa acumulativa. A diferencia del aprendizaje supervisado, que solo considera una sola acción en cada momento, el aprendizaje por refuerzo se ocupa de la secuencia óptima de acciones. Por lo tanto, es una forma de programación dinámica que se utiliza para las decisiones que conducen a una ejecución óptima de operaciones, asignación de cartera y liquidación en un horizonte determinado. Por ejemplo: considere el ejemplo sacado de <https://towardsdatascience.com/applications-of-reinforcement-learning-in-real-world-1a94955bcd12>: Imagine que a un bebé se le da un control remoto de TV en su hogar (entorno). En términos simples, el bebé (agente) primero observará y construirá su propia representación del entorno (estado). Entonces el bebé curioso realizará determinadas acciones como golpear el mando a distancia (acción) y observar cómo sería la respuesta del televisor (siguiente estado). Como un televisor que no responde es aburrido, al bebé no le gusta (recibe una recompensa negativa) y tomará menos acciones que conducirán a tal resultado (actualizar la política) y viceversa. El bebé repetirá el proceso hasta que encuentre una política (qué hacer en diferentes circunstancias) con la que esté satisfecho (maximizando las recompensas totales (descontadas)). En éste curso nos encargaremos del análisis de series de tiempo desde el enfoque del aprendizaje supervisado, en el cual el problema fundamental es el de predicción, es decir, construir un predictor no-lineal, $\hat{Y}(X)$, de una salida Y , dado una matriz(en éste caso un vector) de entrada altamente dimensional $X = (X_1, \dots, X_p)$, donde p es el número de variables. El aprendizaje supervisado usa un modelo parametrizado(paramétrico) $g(X|\theta)$ sobre las covariables o variables explicativas, para predecir la salida Y , la cual puede ser una variable continua o una variable discreta. θ es conocido como el parámetro. Vale la pena decir, que se entenderá por modelo no-paramétrico si el espacio de parámetros es infinito dimensional, mientras que se llama paramétrico si el espacio paramétrico es finito dimensional.

Las redes neuronales representan un mapeo(función) no-lineal $F(X)$ sobre una entrada que toma valores sobre un espacio altamente dimensional usando capas jerárquicas de abstracciones. La teoría de las redes neuronales están basadas en el teorema de representación de Kolmogorov-Arnold de funciones multivariadas(campos escalares). En 1989, [17] demuestra que con una capa oculta, las redes neuronales son aproximadores universales de funciones no-lineales. Una versión mas fuerte del resultado anterior, consiste

que cualquier función continua $f : [0, 1]^n \rightarrow \mathbb{R}^m$ puede ser aproximada a cualquier grado de precisión por una red neuronal feedforward con n nodos de entrada, $2n + 1$ unidades ocultas, m puertas de salida y tres capas.

El aprendizaje automático supervisado es a menudo una forma algorítmica de estimación de modelos estadísticos en la que el proceso de generación de datos se trata como algo desconocido. La selección e inferencia de modelos está automatizada, con énfasis en el procesamiento de grandes cantidades de datos para desarrollar modelos robustos. Puede verse como una técnica de compresión de datos altamente eficiente diseñada para proporcionar predictores en entornos complejos donde las relaciones entre las variables de entrada y salida no son lineales y el espacio de entrada suele ser de gran dimensión. Los algoritmos de aprendizaje automático equilibran el filtrado de datos con el objetivo de tomar decisiones precisas y sólidas, a menudo discretas y como una función categórica de los datos de entrada.

Esto difiere fundamentalmente de los estimadores de máxima verosimilitud utilizados en los modelos estadísticos estándar, que asumen que los datos fueron generados por el modelo y, por lo general, tienen dificultades con el ajuste excesivo, especialmente cuando se aplican a conjuntos de datos de alta dimensión. Dada la complejidad de los conjuntos de datos modernos, como es el caso de series de tiempo financieras de alta dimensión, es cada vez más cuestionable si podemos postular inferencias sobre la base de un proceso de generación de datos conocido. Es una afirmación razonable, incluso si se puede dar una interpretación económica del proceso de generación de datos, que la forma exacta no se puede conocer todo el tiempo.

Por lo tanto, el paradigma que proporciona el aprendizaje automático para el análisis de datos es muy diferente del marco de prueba y modelado estadístico tradicional. Las métricas de ajuste tradicionales, como R^2 , valores t, valores p y la noción de significación estadística, se reemplazan por pronósticos fuera de la muestra y comprensión de la compensación entre sesgo y varianza.

7.7.5 Redes Neuronales Multicapa

Vamos por ahora a suponer por ahora, que las observaciones son independientes e idénticamente distribuidas. Las redes neuronales feedforward son una forma del aprendizaje supervisado que usa capas jerárquicas de abstracción para representar predictores no-lineales de alta dimensionalidad. Recorde-

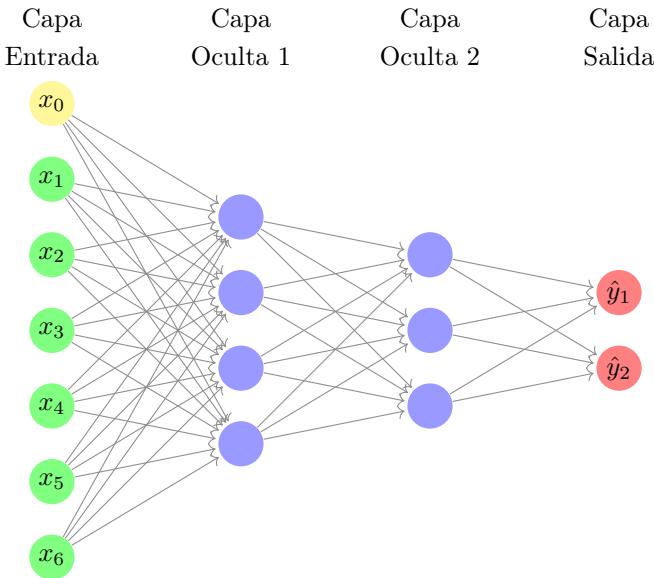


Figura 7.8: Red Neuronal Feedforward

mos que el aprendizaje profundo es manejado por los datos y se enfoca en encontrar estructura en grandes conjuntos de datos. Un modelo de red neuronal feedforward toma la forma general de un mapeo parametrizado

$$Y = F_{W,b}(X) + \varepsilon \quad (7.24)$$

donde $F_{W,b}(X)$ es una red neuronal con L capas y ε es un error i.i.d, ver figura 7.8. Vamos a considerar que la verdadera función que generó los datos fue $F(\cdot)$ y la vamos a representar o aproximar por medio de la red neuronal $F_{W,b}(X)$.

Note que la red neuronal se puede ver como la composición de funciones mas simples:

$$F_{W,b}(X) = f_{W^{(L)}, b^{(L)}}^{(L)} \circ \cdots \circ f_{W^{(1)}, b^{(1)}}^{(1)}(X), \quad (7.25)$$

donde $W = (W^{(1)}, \dots, W^{(L)})$ son las matrices de pesos y $b = (b^{(1)}, \dots, b^{(L)})$ son los vectores de sesgos. Con esto, una vez estimemos los parámetros del modelo, podemos usar la red neuronal para obtener la predicción:

$$\hat{Y}(X) := F_{\hat{W}, \hat{b}}(X) = f_{\hat{W}^{(L)}, \hat{b}^{(L)}}^{(L)} \circ \cdots \circ f_{\hat{W}^{(1)}, \hat{b}^{(1)}}^{(1)}(X), \quad (7.26)$$

Cualquier matriz $W^{(\ell)} \in \mathcal{R}^{m \times n}$ es conformado por vectores columna tal que $W^{(\ell)} = [\mathbf{w}_{,1}^{(\ell)}, \dots, \mathbf{w}_{,n}^{(\ell)}]$ y cada entrada de la matriz $W^{(\ell)}$ se denotará

como $w_{ij}^{(\ell)} = [W^{(\ell)}]_{ij}$.

El mapeo anterior puede verse como una composición de funciones semi-afine.

Definición 7.42. *Sea $\sigma : \mathbb{R} \rightarrow B \subset \mathbb{R}$ una función continua y monótonamente creciente cuyo codominio está en un subconjunto acotado de la recta real. Una función $f_{W^{(\ell)}, b^{(\ell)}}^{(\ell)} : \mathbb{R}^n \rightarrow \mathbb{R}^m$ dada por $f(v) = W^{(\ell)}\sigma^{(\ell-1)}(v) + b^{(\ell)}$, es una función semi-afine en v , por ejemplo, $f(v) = wtanh(v) + b$. σ son las funciones de activación de la salida de la capa anterior.*

Nota 7.43. ✓ Note que si todas las funciones de activación son lineales, entonces $F_{W,b}$ representa justamente una regresión lineal. Contrariamente, si la función de activación es no-lineal, entonces podemos introducir no-linealidades en el modelo, en particular entre los términos de las entradas X_iX_j .

✓ Otras funciones de activación usadas son:

✗ Rectified Linear Unit

$$\sigma(v) = \max\{0, v\}.$$

✗ Sigmoide o softmax

$$\sigma(v) = \frac{1}{1 + e^{-v}}$$

✗ Leaky Relu

$$\sigma(v) = \max\{\alpha v, v\}$$

Teorema de representación Universal

el sustento matemático fundamental en redes neuronales es el teorema de representación universal. Básicamente, en [17] se demuestra que una red neuronal con una sola capa oculta puede aproximar cualquier función continua, sin importar la elección de la función de activación o los datos. Más específicamente, sea $C^p = \{F : \mathbb{R}^p \rightarrow \mathbb{R} | F(x) \in C(\mathbb{R})\}$ el conjunto de las funciones continuas de \mathbb{R}^p a \mathbb{R} . Sea $\Sigma^p(g)$ la clase de funciones $\{F : \mathbb{R}^p \rightarrow \mathbb{R} | F(x) = W^{(2)}\sigma(W^{(1)}x + b^{(1)}) + b^{(2)}\}$. Ahora considere $\Omega = (0, 1]$ y sea C_0 la colección de todos los intervalos abiertos en Ω . Entonces $\sigma(C_0)$ es la σ -álgebra de Borel generada por C_0 , y la denotaremos por $\mathcal{B}((0, 1])$. Sea $M^p = \{F : \mathbb{R}^p \rightarrow \mathbb{R} | F(x) \in \mathcal{B}(\mathbb{R})\}$ el conjunto de todas las funciones Borel medible de \mathbb{R}^p a \mathbb{R} y la σ -álgebra de Borel de \mathbb{R}^p a \mathbb{R} se denotará como \mathcal{B}^p . Entonces el teorema de representación universal es como sigue:

Teorema 7.44. *Teorema de Representación Universal [17]*

Para cada función de activación creciente σ , cada entrada de dimensión p , y cada medida de probabilidad μ sobre $(\mathbb{R}^p, \mathcal{B}^p)$, $\Sigma^p(g)$ es uniformemente denso sobre los compactos en C^p y ρ_μ -denso en M^p .

Este teorema muestra que las redes feedforward estándar con una sola capa oculta pueden aproximar arbitrariamente bien cualquier función uniformemente continua en cualquier conjunto compacto y cualquier función medible en la métrica ρ_μ , independientemente de la función de activación (siempre que sea medible), independientemente de la dimensión del espacio de entrada, p , e independientemente del espacio de entrada. En otras palabras, tomando el número de unidades ocultas, k , lo suficientemente grande, cada función continua sobre \mathbb{R}^p puede aproximarse con bastante precisión, de manera uniforme sobre cualquier conjunto acotado por funciones obtenidas de las redes neuronales con una capa oculta.

Nota 7.45. ✓ El teorema de aproximación caracteriza a las redes neuronales feedforward con una sola capa oculta como una familia de soluciones aproximadas.

- ✓ Note que el teorema no especifica como configurar un perceptrón multicapa tal que nos informe acerca de las propiedades de la aproximación.
- ✓ El teorema no dice nada al respecto de añadir mas capa ocultas, es decir, no nos indica que tan bueno o malo, es aumentar el número de capas.
- ✓ Con el objeto de estimar los parámetros W, b de la red neuronal, el teorema no nos indica si los métodos numéricos (como el descenso del gradiente), pueden optimizar en finitos pasos, la función objetivo.
- ✓ El teorema tampoco caracteriza el error de predicción de ninguna manera, el cual es vital debido al problema del exceso de ajuste, para datos fuera de la muestra.
- ✓ El teorema tampoco informa acerca de como las redes neuronales multicapa pueden cubrir otras técnicas de aproximación como casos especiales, tales como, interpolación spline polinomial.
- ✓ La razón de por qué múltiples capas ocultas son necesarias es un problema abierto aún. Sin embargo varias caracterizaciones se han dado, ver [6] Páginas 127-132 para caracterizaciones usando la dimensión

VC, [32] Páginas 243-250 usando el enfoque de cotas para el ECM de la función de “Regresión” .

- ✓ *Un teorema de aproximación universal general y su respectiva caracterización puede ser encontrado en [23].*

Nota 7.46. Vale la pena decir que existen errores cuando se utilizan las redes neuronales para aproximar una función. Además del poder expresivo, que determina el error de aproximación del modelo, existe la noción de capacidad de aprendizaje, que determina el nivel de error de estimación. El primero mide el error introducido por una función de aproximación y el segundo mide el rendimiento perdido como resultado de utilizar una muestra de entrenamiento finita.

Una medida clásica de capacidad de aprendizaje es dada por la dimensión Vapnik-Chervonenkis(VC). La dimensión VC determina las condiciones necesarias y suficientes para la consistencia y la tasa de convergencia de los procesos de aprendizaje (es decir, el proceso de elegir una función apropiada de un conjunto dado de funciones). Si una clase de funciones tiene una dimensión VC finita, entonces se puede aprender. Esta medida de capacidad es más robusta que medidas arbitrarias como el número de parámetros. Es posible, por ejemplo, encontrar un conjunto simple de funciones que dependa de un solo parámetro y que tenga una dimensión de VC infinita.

Vapnik en su libro [39] formuló un método de inferencia inductiva basada en la dimensión VC. Este enfoque, conocido como minimización del riesgo empírico estructural, logró el límite más pequeño en el error de prueba utilizando los errores de entrenamiento y eligiendo la máquina (es decir, el conjunto o las funciones) con la dimensión de VC más pequeña. El problema de minimización expresa el equilibrio entre sesgo y varianza. Por un lado, para minimizar el sesgo, es necesario elegir una función de un amplio conjunto de funciones, no necesariamente con una dimensión de VC baja. Por otro lado, la diferencia entre el error de entrenamiento y el error de prueba (es decir, la varianza) aumenta con la dimensión de VC (también conocida como expresibilidad o generalización).

Note que el error cuadrático medio(MSE) entre \hat{F} y Y o el error de predicción esperado puede escribirse como sigue para un valor fijo de X (por lo

general para un valor X que no está en el conjunto de entrenamiento):

$$MSE(Y, \hat{F}_{W,b}) = E[(Y - \hat{F}_{W,b})^2] \quad (7.27)$$

$$= (E[\hat{F}_{W,b}(X)] - F(X))^2 + E[\hat{F}_{W,b}(X) - E(\hat{F}_{W,b}(X))]^2 + \sigma_{\varepsilon}^2 \quad (7.28)$$

$$= Sesgo^2(\hat{F}_{W,b}(X)) + Var(\hat{F}_{W,b}(X)) + \sigma_{\varepsilon}^2 \quad (7.29)$$

$$= \text{Error Irreducible} + Sesgo^2 + Varianza. \quad (7.30)$$

Observe que el MSE hace referencia a el promedio de error que uno debería obtener si estimamos repetidamente a F usando un gran número de muestras de entrenamiento, y lo probamos cada vez en X . La relación entre sesgo y varianza está estrechamente ligada a los conceptos de capacidad de aprendizaje automático, subajuste y sobreajuste. Cuando el error de generalización es medido por el MSE (donde el sesgo y la varianza son componentes significativos del error de generalización), el aumento de la capacidad tiende a aumentar la varianza y disminuir los sesgos. Recuerde que la verdadera función es F y nosotros la estamos aproximando a través de la red neuronal $\hat{F}_{W,b}$ una vez se han estimado los parámetros W, b . Lo deseable sería minimizar $MSE(Y, \hat{F}_{W,b})$, y eso se hace minimizando el sesgo y la varianza al mismo tiempo, ya que σ_{ε}^2 es irreducible. Sin embargo, en la práctica no es posible minimizar ambos al mismo tiempo, por lo tanto se requiere un compromiso(compensación) entre el sesgo y la varianza, es decir, es fácil obtener un método con un sesgo extremadamente bajo pero una alta varianza (por ejemplo, dibujando una curva que pase por cada observación de entrenamiento) o un método con una varianza muy baja pero con un alto sesgo (ajustando una línea horizontal a los datos), ver libros [14] y [20]. Como regla general se tiene que si se usan métodos mas flexibles, la varianza va a incrementarse mientras que el sesgo se irá a disminuir.

Lo anterior se puede ver en general en términos de riesgo. El riesgo esperado es una medida de rendimiento fuera de la muestra del modelo aprendido y se basa en la función de densidad de probabilidad conjunta (pdf) $p(x, y)$:

$$R[\hat{F}] = \mathbb{E}[\mathcal{L}(\hat{F}(X), Y)] = \int \mathcal{L}(\hat{F}(X), Y) dp(x, y). \quad (7.31)$$

Note que si se pudiera elegir una función \hat{F} que minimice el riesgo esperado, entonces uno tendría una medida del optimo aprendizaje. Note sin embargo que el riesgo esperado no se puede medir directamente ya que la función de densidad es desconocida. Entonces una estimación de esa cantidad es requerida. El estimador de ese riesgo esperado se conoce como medida

de riesgo empírico(ERM):

$$R_{emp}(\hat{F}) = \frac{1}{N} \sum_{i=1}^N \mathcal{L}(\hat{F}(\mathbf{x}_i), \mathbf{y}_i). \quad (7.32)$$

Note que por la ley de los grandes números bajo el supuesto i.i.d, el riesgo empírico converge al riesgo esperado. Se puede verificar que la dimensión VC provee bandas para el riesgo esperado en función del ERM y el numero de observaciones de entrenamiento:

$$R[\hat{F}] \leq R_{emp}(\hat{F}) + \sqrt{\frac{h(\ln(2N/h) + 1) - \ln(n/4)}{N}} \quad (7.33)$$

donde h es la dimensión VC de $\hat{F}(X)$ y $N > h$. La figura 7.9 , extraída del libro [6] muestra la compensación entre la dimensión VC y la rigidez de la banda.

Fig. 4.6 This figure shows the tradeoff between VC dimension and the tightness of the bound. As the ratio N/h gets larger, i.e. for a fixed N , we decrease h , the VC confidence becomes smaller, and the actual risk becomes closer to the empirical risk. On the other hand, choosing a model with a higher VC dimension reduces the ERM at the expense of increasing the VC confidence

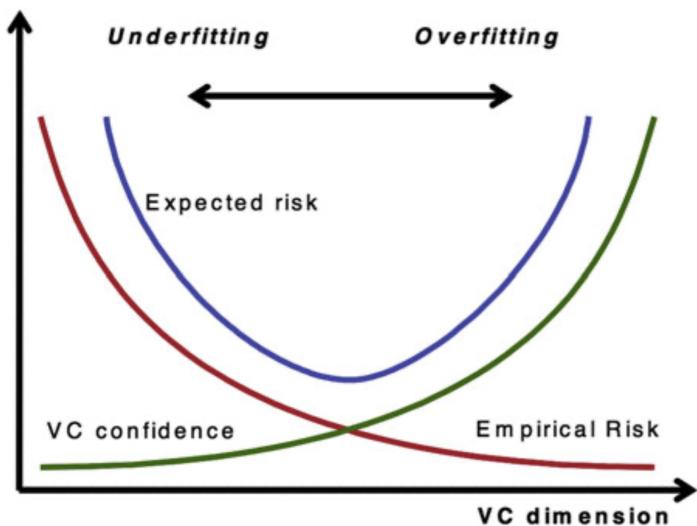


Figura 7.9:

Nota 7.47. Por lo tanto, bajo una configuración especial de los pesos y sesgos, con las unidades ocultas que definen las celdas de Voronoi(El diagrama de Voronoi es una división de un plano en regiones según la distancia a los puntos en un subconjunto específico del plano.) para cada observación, podemos mostrar que una red neuronal es una spline univariante. Este resultado se generaliza a splines de mayor dimensión y orden superior. Tal

resultado nos permite ver las splines como un caso especial de una red neuronal que es consistente con nuestro razonamiento de las redes neuronales como técnicas de aproximación y regresión generalizadas. La formulación de redes neuronales como splines permite que la teoría de la aproximación oriente el diseño de la red. Desafortunadamente, equiparar redes neuronales con splines dice poco sobre por qué y cuándo se necesitan múltiples capas.

Se ha demostrado que las redes profundas pueden lograr un rendimiento superior en comparación con los modelos aditivos lineales, como la regresión lineal, evitando al mismo tiempo la maldición de la dimensionalidad. Además, existen muchos desarrollos teóricos recientes que caracterizan el comportamiento de aproximación en función de la profundidad, el ancho y el nivel de dispersión de la red. También se ha probado la importancia de múltiples capas para la red, la cual considera el efecto de componer funciones afines a trozos en lugar de agregarlas.

Finalmente algunas condiciones deben ser impuestas para restringir el rango de $\hat{f}(c)$ o imponer ciertas propiedades acerca de la forma de la función $f(x)$ que será aproximada. Estas condiciones tiene que ver con la convexidad de las funciones de activación y las restricciones acerca de los pesos. Para mas detalles ver Página 133 del libro [6].

Vale la pena decir que hay otras alternativas similares a las MLP, por ejemplo, la tabla 7.10, sacada del libro [6] nos los muestra:

Entrenamiento, Validación y Prueba

El aprendizaje profundo es un enfoque basado en datos que se centra en encontrar estructuras en grandes conjuntos de datos. Las principales herramientas para la selección de variables o predictores son la regularización y el abandono. El rendimiento predictivo fuera de la muestra ayuda a evaluar la cantidad óptima de regularización, es decir, problema de encontrar la selección óptima de hiperparámetros. El procedimiento de modelado y el modelador sigue dos pasos clave:

Fase de entrenamiento: Empareje la entrada(y_i) con la salida esperada(\hat{y}_i), hasta que se haya encontrado una coincidencia lo suficientemente cercana.

Function class \mathcal{F} (and its parameterization)	Regularizer $R(f)$
<i>Global/parametric predictors</i>	
Linear $\beta'x$ (and generalizations)	Subset selection $\ \beta\ _0 = \sum_{j=1}^k \mathbf{1}_{\beta_j \neq 0}$
	LASSO $\ \beta\ _1 = \sum_{j=1}^k \beta_j $
	Ridge $\ \beta\ _2^2 = \sum_{j=1}^k \beta_j^2$
	Elastic net $\alpha \ \beta\ _1 + (1 - \alpha) \ \beta\ _2^2$
<i>Local/non-parametric predictors</i>	
Decision/regression trees	Depth, number of nodes/leaves, minimal leaf size, information gain at splits
Random forest (linear combination of trees)	Number of trees, number of variables used in each tree, size of bootstrap sample, complexity of trees (see above)
Nearest neighbors	Number of neighbors
Kernel regression	Kernel bandwidth
<i>Mixed predictors</i>	
Deep learning, neural nets, convolutional neural networks	Number of levels, number of neurons per level, connectivity between neurons
Splines	Number of knots, order
<i>Combined predictors</i>	
Bagging: unweighted average of predictors from bootstrap draws	Number of draws, size of bootstrap samples (and individual regularization parameters)
Boosting: linear combination of predictions of residual	Learning rate, number of iterations (and individual regularization parameters)
Ensemble: weighted combination of different predictors	Ensemble weights (and individual regularization parameters)

Figura 7.10: Alternativas similares a MLP

Fase de validación y prueba: Evalúe qué tan bien se ha entrenado al aprendiz profundo para la predicción fuera de la muestra. Esto depende del tamaño de sus datos, el valor que le gustaría predecir, la entrada, etc., y varias propiedades del modelo, incluido el error medio para los predictores numéricos y los errores de clasificación para los clasificadores.

La fase de validación y prueba es subdividida en dos partes:

- 2.a. Primero estime la precisión fuera de la muestra para todos los enfoques(modelos)(*validación*). Note que con el conjunto de validación también se podrán encontrar los hiperparámetros del modelo tal que el modelo no presenten problemas de sobre ajuste.
- 2.b. Compare los modelos y el seleccione el enfoque de mejor rendimiento basado en los datos de validación(*verificación*). Podría usarse para ahora si, decir el modelo final.

Nota 7.48. *El paso 2.b. puede saltarse si no hay necesidad de seleccionar un modelo de un conjunto de enfoques(o aproximaciones).*

Por otro lado, para construir y evaluar un modelo de aprendizaje automático, vamos a considerar que tenemos unos datos de entrenamiento $D = \{Y^{(i)}, X^{(i)}\}_{i=1}^N$, es decir los datos para estimar los parámetros del modelo. Con el objeto de encontrar $Y = F(X)$, es necesario establecer una función de pérdida para un predictor \hat{Y} de la variable de salida Y , es decir $\mathcal{L}(Y, \hat{Y})$. Si suponemos que existe un modelo de probabilidad subyacente, $p(Y|\hat{Y})$, entonces la función de pérdida es el negativo del log de la función de probabilidad, es decir, $\mathcal{L}(Y, \hat{Y}) = -\log p(Y|\hat{Y})$. Por ejemplo, si asumimos que el modelo es Gaussiano, entonces $\mathcal{L}(Y, \hat{Y}) = ||Y - \hat{Y}||^2$ es la norma L^2 o la pérdida cuadrática; para clasificación binaria tenemos $\mathcal{L}(Y, \hat{Y}) = -Y \log \hat{Y}$ que es la entropía cruzada negativa, proveniente de un modelo Bernoulli. En su forma mas simple, el problema de optimización consiste de

$$\underset{W,b}{\text{minimizar}} f(W, b) + \lambda \phi(W, b)$$

tal que

$$f(W, b) = \frac{1}{N} \sum_{i=1}^N \mathcal{L}(Y^{(i)}, \hat{F}(X^{(i)})) = \frac{1}{N} \sum_{i=1}^N \mathcal{L}(Y^{(i)}, \hat{Y}(X^{(i)}))$$

con un componente de penalización por regularización $\phi(W, b)$. La regularización es cualquier modificación que hagamos en un algoritmo de aprendizaje que tenga como objetivo reducir su error de generalización pero no su

error de entrenamiento. Note que la ecuación anterior, es equivalente a minimizar el riesgo empírico definido en 7.32. En general la función de pérdida es no convexa, lo cual hace que encontrar un mínimo local sea difícil debido a que puede poseer varios mínimos locales. Un supuesto que debe imponer a los errores es que tenga la misma distribución, sin embargo éste supuesto es relajado ponderando las observaciones de forma diferente. Vale la pena decir que se considera tales extensiones tan sencillo y compatible con los algoritmos para resolver el problema no ponderado problema de optimización. Observe que λ es el parámetro de regularización global el cual se afina usando el error cuadrático medio predictivo fuera de la muestra(de entrenamiento) del modelo. Note también que la penalidad por regularización, $\phi(W, b)$ induce una compensación entre el sesgo y la varianza. Hay muchas formas de incluir la penalización en el modelo, por ejemplo, $\phi(W, b)$ puede ser una norma. Así λ penaliza la norma. Ejemplos de la norma pueden ser : norma L^2 , es decir, $\phi(\theta) = \frac{1}{2} \|\theta\|_2^2$. Norma L^1 , es decir, $\phi(\theta) = \|\theta\|_1 = \sum |\theta_i|$. También se puede incluir una condición que permita condicionar a que la penalización $\phi(\theta)$ sea mas pequeño que un valor k , es decir, ahora la la penalización queda

$$\lambda(\phi(W, b) - k).$$

Existen otras estrategias que ayudar a evitar el sobre ajuste como por ejemplo: "detener temprano"(early stopping), data augmentation entre otros. Para mas información, ver [13] capítulo 7.

Finalmente, la optimización(minimización) se hace a través de un método numérico el cual es llamado el algoritmo del descenso del gradiente, el cual requiere del computo del gradiente $\nabla \mathcal{L}$ en cada iteración a través de lo que se conoce como el algoritmo back-propagation. En seguida veremos en que consiste el método del descenso del gradiente.

El Descenso del Gradiente Estocástico(SGD)

El algoritmo del descenso del gradiente o cualquiera de sus variaciones, es usado para encontrar(estimar) los pesos y los interceptos(sesgos)de un modelo de aprendizaje profundo, minimizando la función de pérdida penalizada $f(W, b)$. La idea principal del método es que minimiza la función de pérdida dando un paso negativo a lo largo de una estimación g^k del gradiente $\nabla f(W^k, b^k)$ en la iteración k . La estimación(aproximación) del gradiente es calculado por

$$g^k = \frac{1}{b^k} \sum_{i \in E_k} \nabla \mathcal{L}_{W,b}(Y^{(i)}, \hat{Y}^k(X^{(i)}))$$

donde $E_k \subset \{1, \dots, N\}$ y $b_k = |E_k|$ es el cardinal de E_k , el cual se conoce como el tamaño del lote. Cuando $b_k > 1$ el algoritmo es llamado el descenso del gradiente por lotes(SGD por lotes) o simplemente SGD. Una estrategia usual para elegir el subconjunto E es ir cíclicamente y tomar elementos consecutivos de $\{1, \dots, N\}$ y $E_{k+1} = [E_k \mod N] + 1$. La dirección aproximada g^k es calculada usando la regla de la cadena, es decir, *back-propagation* para el aprendizaje profundo. Se puede verificar que g^k es un estimador insesgado de $\nabla f(W^k, b^k)$. En cada iteración, nosotros actualizamos la solución $(W, b)^{k+1} = (W, b)^k - t_k g^k$.

En aplicaciones de aprendizaje profundo, se usa un tamaño de paso t_k , mejor conocido como tasa de aprendizaje, como una constante o una estrategia de reducción de la forma $t_k = a \exp\{-kt\}$.

Nota 7.49. Consideré que tenemos los datos del aprendizaje supervisado de la siguiente manera:

$$y_1 \quad \underline{x}_1$$

$$y_2 \quad \underline{x}_2$$

$$\vdots \quad \vdots$$

$$y_{20} \quad \underline{x}_{20}$$

y consideramos lotes a manera de ejemplo de tamaño $l = 5$, entonces tendremos 4 lotes para la muestra de tamaño $n = 20$ y cada lote estará configurado como sigue:

<i>Lote 1</i>	y_1	\underline{x}_1
	y_2	\underline{x}_2
	\vdots	\vdots
	y_5	\underline{x}_5
<hr/>		
<i>Lote 2</i>	y_6	\underline{x}_6
	y_7	\underline{x}_7
	\vdots	\vdots
	y_{10}	\underline{x}_{10}
<hr/>		
<i>Lote 3</i>	y_{11}	\underline{x}_{11}
	y_{12}	\underline{x}_{12}
	\vdots	\vdots
	y_{15}	\underline{x}_{15}
<hr/>		
<i>Lote 4</i>	y_{16}	\underline{x}_{16}
	y_{17}	\underline{x}_{17}
	\vdots	\vdots
	y_{20}	\underline{x}_{20}

Back-Propagation

Para motivar el método de Back-Propagation, usaremos el ejemplo de multi-clasificación. Consideremos que el espacio de estado de la predicción $\hat{Y} \in [0, 1]^k$ como una función de la matriz final de pesos $W \in \mathbb{R}^{K \times M}$ y el sesgo de salida $b \in \mathbb{R}^K$, tal que

$$\hat{Y}(W, b) = \sigma \circ I(W, b), \quad (7.34)$$

donde la función $I : \mathbb{R}^{K \times M} \times \mathbb{R}^K \rightarrow \mathbb{R}^K$, es de la forma $I(W, b) := WX + b$ y $\sigma : \mathbb{R} \rightarrow \mathbb{R}$ es la función softmax. Note que si estamos en un problema de

regresión, es decir el espacio de estado es \mathbb{R} o \mathbb{R}^K para la variable de salida Y , entonces la función de salida se modifica a que sea lineal, o cualquier otra dependiendo el espacio de estado de dicha variable. Aplicando la regla de la cadena multivariada a $\hat{Y}(W, b)$, tenemos el jacobiano, el cual es de la siguiente forma:

$$\nabla \hat{Y}(W, b) = \nabla(\sigma \circ I)(W, b) \quad (7.35)$$

$$= \nabla \sigma(I(W, b)) \cdot \nabla I(W, b). \quad (7.36)$$

Actualización de los pesos

Recordemos que la función de pérdida para un multi-clasificador es la entropía cruzada

$$\mathcal{L}(Y, \hat{Y}(X)) = - \sum_{k=1}^K Y_k \ln \hat{Y}_k. \quad (7.37)$$

note que si estamos encaso de regresión, podemos usar la pérdida cuadrática $\|Y - \hat{Y}\|^2$.

Ahora, puesto que Y es un vector de constantes, es posible expresar la entropía cruzada como una función de (W, b)

$$\mathcal{L}(W, b) = \mathcal{L} \circ \sigma(I(W, b)). \quad (7.38)$$

Aplicando la regla de la cadena multivariada nos da:

$$\nabla \mathcal{L}(W, b) = \nabla(\mathcal{L} \circ \sigma)(I(W, b)) \quad (7.39)$$

$$= \nabla \mathcal{L}(\sigma(I(W, b))) \cdot \nabla \sigma(I(W, b)) \cdot \nabla I(W, b). \quad (7.40)$$

El algoritmo del descenso del gradiente estocástico es usado para encontrar el mínimo

$$(\hat{W}, \hat{b}) = \arg \min_{W, b} \frac{1}{N} \mathcal{L}(y_i, \hat{Y}^{W, b}(\mathbf{x}_i)). \quad (7.41)$$

Note que el modelo es formado a través de composiciones de funciones simples, entonces el gradiente debe obtenido usando la regla de la cadena obtener su derivada. Este cómputo es hecho en dos pasos de "barrido", uno hacia delante y otro hacia atrás(back-propagation) sobre toda la red neuronal, haciendo un seguimiento sólo de las cantidades locales a cada neurona.

Paso hacia adelante(Forward)

Sea $Z^{(0)} = X$ y para $\ell \in \{1, \dots, L\}$ establezcamos

$$Z^{(\ell)} = f_{W^{(\ell)}, b^{(\ell)}}^{(\ell)}(Z^{(\ell)}) = \sigma^{(\ell)}(W^{(\ell)}Z^{(\ell-1)} + b^{(\ell)}). \quad (7.42)$$

Con un paso completo hacia adelante, el error $\hat{Y} - Y$ es evaluado usando $\hat{Y} = Z^{(L)}$.

Paso hacia atrás(Back-Propagation)

Definamos el error propagación hacia atrás $\delta^{(\ell)} = \nabla_{b^{(\ell)}} \mathcal{L}$, dado por $\delta^{(L)} = \hat{Y} - Y$, y para $\ell = L-1, \dots, 1$ la siguiente relación de recurrencia nos da la actualización del error de propagación hacia atrás y de los pesos para la capa ℓ :

$$\delta^{(\ell)} = (\nabla_{I^{(\ell)}} \sigma^{(\ell)}) W^{(\ell+1)'} \delta^{(\ell+1)}, \quad (7.43)$$

$$\nabla_{W^{(\ell)}} \mathcal{L} = \delta^{(\ell)} \otimes Z^{(\ell-1)}. \quad (7.44)$$

El símbolo \otimes quiere decir el producto externo de dos vectores el cual es una matriz. Los pesos y los sesgos son actualizados para todo $\ell \in \{1, 2, \dots, L\}$ de acuerdo a la expresión

$$\Delta W^{(\ell)} = -\gamma \nabla_{W^{(\ell)}} \mathcal{L} = -\gamma \delta^{(\ell)} \otimes Z^{(\ell-1)} \quad (7.45)$$

$$\Delta b^{(\ell)} = -\gamma \delta^{(\ell)}, \quad (7.46)$$

donde γ es la tasa de aprendizaje definida por el usuario. Vale la pena notar que el signo negativo indica que los pesos cambian en la dirección en que se disminuye el error. Las actualizaciones por mini-lotes o fuera de línea implican el uso de muchas observaciones de X al mismo tiempo. El tamaño del lote se refiere al número de observaciones de X utilizadas en cada pasada. Una época se refiere a un viaje de ida y vuelta (es decir, pase hacia adelante + hacia atrás) sobre todas las muestras de entrenamiento.

Momentum

Una de las desventajas del algoritmo del descenso del gradiente es que f puede no ser alcanzada o cada actualización puede ser lenta. Adicionalmente, la varianza de la estimación del gradiente g^k es cercana a cero a medida que en las iteraciones van convergiendo a la solución. Una solución para esto consiste en unir dos conceptos: descenso por "coordinada" y modificaciones basadas en "momentum". La idea del descenso por coordinada es que en cada

7.8. REDES NEURONALES PARA MÚLTIPLE-RESPUESTA DE RESPUESTA VECTORIAL

paso se evalúa sólo una componente E_k del gradiente ∇f , mientras que el concepto de momentum quiere decir el gradiente sólo influencia directamente los cambios en la velocidad de la actualización(e.d. en la tasa de aprendizaje), mas específicamente:

$$\begin{aligned} v^{k+1} &= \mu v^k - t_k g^k \\ (W, b)^{k+1} &= (W, b)^k + v^k. \end{aligned}$$

μ controla el efecto de descarga sobre la tasa de aprendizaje de las variables. El nombre momentum viene de la analogía con la reducción en la energía cinética que permite ir mas despacio en lo movimiento hacia el mínimo.

Otra modificación al método SGD es el método “AdaGrad”, el cual escala en forma adaptativa cada uno de los parámetros de aprendizaje en cada iteración:

$$\begin{aligned} c^{k+1} &= c^k + g((W, b)^k)^2 \\ (W, b)^{k+1} &= (W, b)^k - t_k g((W, b)^k)(\sqrt{c^{k+1}} - a) \end{aligned}$$

donde a es un número pequeño que evita la división por cero. Por otro lado, PRMSprop toma la idea de AdaGrad más allá y pone más peso en los valores recientes del gradiente al cuadrado para escalar la dirección de actualización, es decir, tenemos:

$$c^{k+1} = dc^k + (1 - d)g((W, b)^k)^2.$$

El método *Adam* combina los métodos PRMSprop y momentum condice a la siguientes ecuaciones de actualización:

$$\begin{aligned} v^{k+1} &= \mu v^k - (1 - \mu)t_k g((W, b)^k + v^k), \\ c^{k+1} &= dc^k + (1 - d)g((W, b)^k)^2, \\ (W, b)^{k+1} &= (W, b)^k - t_k v^{k+1}/(\sqrt{c^{k+1}} - a). \end{aligned}$$

7.8 Redes Neuronales para Múltiple-Respuesta de respuesta Vectorial

Red Neuronal de Una Capa

Esta sección es tomada del libro [30].

Vamos a explicar como es la configuración de datos para entrenar un modelo de redes neuronales cuando se tiene múltiples entradas y múltiples salidas. Inicialmente consideraremos que tenemos una red neuronal con una sola capa oculta, la cual tiene P entradas, M nodos o neuronas, y K nodos de salida. Sea $\underline{x}_t = (x_{1t}, \dots, x_{Pt})'$ el vector de entrada, y $\underline{y}_t = (y_{1t}, \dots, y_{Kt})'$ el vector de salida, y el vector de salida de la capa oculta como $\underline{h}_t = (h_{1t}, \dots, h_{Mt})'$. El caso que vamos a considerar consiste de cuando las respuestas se modelan de forma independiente cada una. El modelo de redes es como sigue:

$$h_{mt} = g(w_{m0} + \underline{w}_m \underline{x}_t), \quad m = 1, \dots, M, \quad (7.47)$$

$$q_{vt} = \beta_{v0} + \underline{\beta}_v \underline{h}_t, \quad v = 1, \dots, K, \quad (7.48)$$

$$y_{vt} = f_v(\underline{x}_t) = g_v(q_t), \quad v = 1, \dots, K, \quad (7.49)$$

donde $g(\cdot)$, $g_v(\cdots)$ son funciones de activación, w_{m0} , $\underline{w}_m = (w_{m1}, \dots, w_{mP})'$ son los sesgos y los pesos asociados con la m -ésima neurona de la capa oculta, β_{v0} , $\underline{\beta}_v = (\beta_{v1}, \dots, \beta_{vM})'$ son respectivamente, el sesgo y los pesos de la v -ésima neurona de salida, y $\underline{q}_t = (q_{1t}, \dots, q_{vM})'$. Esta red se representa gráficamente como sigue:

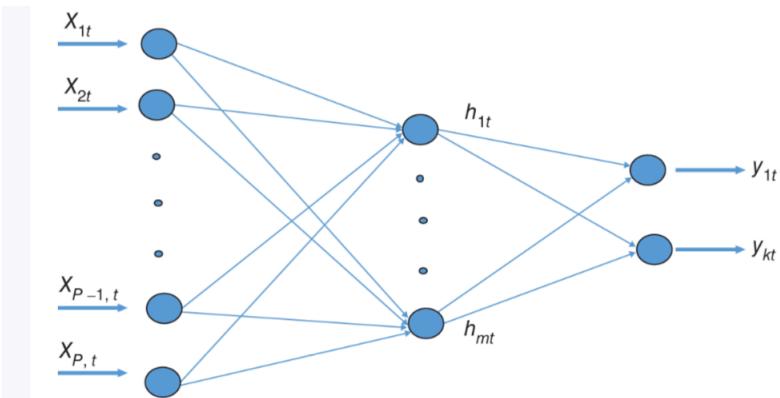


Figura 7.11: Tomado del Libro [30]

El modelo anterior puede verse como un modelo factorial, donde las P variables explicativas son combinadas para crear M combinaciones lineales o factores. Cada factor es entonces transformado en una forma no-lineal para crear las nuevas variables explicativas, h_{mt} , y las K variables de respuesta son funciones de las combinaciones lineales de esas nuevas variables h_{mt} . También, el sesgo puede ser incorporado dentro de los pesos re-definiendo las variables de salida como $\underline{x}_t = (1, x_{1t}, \dots, x_{Pt})'$ y los nodos ocultos como

7.8. REDES NEURONALES PARA MÚLTIPLE-RESPUESTA DE RESPUESTA VECTORIAL

$\underline{h}_t = (1, h_{1t}, \dots, h_{Mt})'$. La función $f_v(\cdot)$ es usado para denotar la función subyacente para la v -ésima variable de salida como función de \underline{x}_t . Para variables continuas y_t , $g_v(\cdot)$ usualmente se asume que sea lineal $g_v(q_t) = q_{vt} = \beta_{v0} + \underline{\beta}'_v \underline{h}_t$.

Nota 7.50. Dado $g(\cdot)$ y $g_v(\cdot)$, la red contiene los siguientes parámetros:

$$\begin{aligned} \{w_{m0}, \underline{w}_m | m = 1, \dots, M\} &: M(P+1) \text{ parámetros} \\ \{\beta_{v0}, \underline{\beta}_v | v = 1, \dots, M\} &: K(M+1) \text{ parámetros}. \end{aligned}$$

Sea Θ el vector de parámetros de la red (sesgos y pesos). Para las variables de respuesta continua, la función de pérdida cuadrática queda dada de la siguiente forma:

$$L(\Theta) = \sum_{v=1}^K \sum_{t=1}^{n_1} [y_{vt} - f_v(\underline{x}_t)]^2. \quad (7.50)$$

donde n_1 es el tamaño de la muestra de entrenamiento.

Nota 7.51. Note que se ha asumido que las observaciones X_i, Y_i para $i = 1, \dots, n$ son i.i.d. Sin embargo en el contexto de series de tiempo, no se puede asumir este supuesto. El artículo [35] muestra que se requiere de algunas condiciones sobre el proceso estocástico $(X_t, Y_t)_{t \geq 1}$, por ejemplo que procesos satisfaga la WLLNE (la ley débil de los grandes números para eventos) para que un método de aprendizaje pueda aprender de forma asintótica (definición 2.16). En particular para SVM, con función de pérdida convexa puede aprender si el proceso estocástico $(X_t, Y_t)_{t \geq 1}$ satisface WLLNE o SLLN.

Adicionalmente, si el proceso estocástico satisface condiciones de mixing, entonces también las SVM pueden aprender.

Red Neuronal Recurrente(RNN)

Una red neuronal recurrente son una familia de redes neuronales para procesar datos secuenciales. Ellas son similares a los sistemas dinámicos

$$\underline{s}_t = f(\underline{s}_{t-1}, \underline{x}_t, \Theta),$$

donde $\underline{s}_t \in \mathbb{R}^d$ denota el vector de estado en el tiempo t , $\underline{x}_t \in \mathbb{R}^k$ es la entrada en el tiempo t , $\Theta \in \mathbb{R}^d$ es el conjunto de parámetros, y $f(\cdot)$ denota la transición del estado \underline{s}_{t-1} al estado \underline{s}_t con los nuevos datos \underline{x}_t . Aquí se asume que el vector de parámetros Θ no varía en el tiempo, es decir,

algunos valores se comparten en todos los tiempos t . Dado los datos de entrada $\{\underline{x}_1, \dots, \underline{x}_n\}$, sea $\underline{h}_t \in \mathbb{R}^d$ la salida de la capa oculta, donde $\underline{h}_0 = \underline{0}$. Entonces una red neuronal recurrente puede ser escrita como sigue:

$$\underline{s}_t = \underline{b} + W\underline{h}_{t-1} + U\underline{x}_t \quad (7.51)$$

$$\underline{h}_t = g(\underline{s}_t), \quad (7.52)$$

$$\hat{\underline{y}}_t = g_y(\underline{c} + V\underline{h}_t), \quad (7.53)$$

donde $g(\cdot)$ y $g_y(\cdot)$ son funciones de activación, $\underline{b}, \underline{c}$ son vectores de sesgos, y W, U y V son los matriciales y $\hat{\underline{y}}_t \in \mathbb{R}^m$. El vector \underline{h}_t se conoce como el vector de memoria. En cada tiempo t , se tiene disponible el vector de entrada \underline{x}_t y el vector de memoria \underline{h}_{t-1} . Esos dos vectores son combinados linealmente para obtener el vector de estado en el tiempo t . Es nuevo vector de memoria es una función no-lineal de el vector de estado, y la salida(pronóstico) es construido como una transformación no-lineal de la combinación lineal de las componentes del nuevo vector de memoria \underline{h}_t . Esto se puede representar gráficamente como sigue:

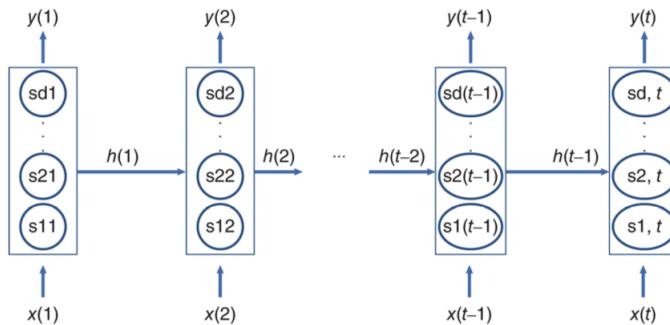


Figura 7.12:

Red LSTM(Long Short-Term Memory)

Usualmente, las redes neuronales recurrentes presentan algunos problemas como es el desvanecimiento o la explosión del gradiente. Las redes neuronales LSTM introducen bucles automáticos que producen trayectorias donde el gradiente puede fluir por largas duraciones. Con esto, se mitiga el desvanecimiento del gradiente pero no su explosión. La red LSTM recurrente usual emplea celdas LSTM que tienen un bucle interno en adición a la usual recurrencia interna. Cada celda LSTM tiene las mismas entradas y salidas de una RNN tradicional, pero emplea tres reguladores, las cuales se les llama

puertas, que gobiernan el flujo de la información dentro de la celda. Consideré la ecuación 7.51 $\underline{s}_t = \underline{b} + W\underline{h}_{t-1} + U\underline{x}_t$. El nodo del estado s_{jt} de \underline{s}_t ahora se acompaña de tres puertas. La primera puerta es llamada la puerta del olvido, y se define como:

$$f_{jt} = \sigma \left(b_j^f + \sum_v W_{jv}^f h_{v,t-1} + \sum_v U_{jv}^f x_{vt} \right), \quad (7.54)$$

donde σ es la función de activación sigmoide. La segunda puerta es llamada la puerta de entrada externa i_{jt} , la cual se define con su propios parámetros

$$i_{jt} = \sigma \left(b_j^i + \sum_v W_{jv}^i h_{v,t-1} + \sum_v U_{jv}^i x_{vt} \right). \quad (7.55)$$

Mientras que la tercera puerta es la puerta de salida, a través de la cual la salida de la memoria h_{it} de la celda LSTM puede ser apagada, y se define como sigue:

$$o_{jt} = \sigma \left(b_j^o + \sum_v W_{jv}^o h_{v,t-1} + \sum_v U_{jv}^o x_{vt} \right). \quad (7.56)$$

Así, la celda de estado es dada por

$$s_{jt} = f_{jt}s_{j,t-1} + i_{jt}\sigma \left(b_j + \sum_v W_{jv} h_{v,t-1} + \sum_v U_{jv} x_{vt} \right),$$

donde la función de activación σ en este caso es tanh. Se puede notar que las puertas del olvida y de entrada externa gobiernan la evolución de los nodos de los estados. Finalmente, tenemos que

$$h_{jt} = \tanh(s_{jt})o_{jt}. \quad (7.57)$$

7.9 Combinación de Pronósticos

Sacado de [29] página 342. Supongamos que se han estimado k modelos posibles para una serie de tiempo y que se dispone de los valores del BIC para cada modelo. Se puede verificar que el BIC es una función de $-2 \ln P(M_i|D)$, donde $P(M_i|D)$ es la probabilidad del modelo i dados los datos D o es la probabilidad aposteriori del modelo. Entonces como

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

entonces tenemos que

$$P(M_i|D) = c_1 e^{-0.5BIC_i}$$

donde $c_1 = \exp(c)$. Para calcular la constante c_1 , se hace el uso de que la suma de las probabilidades de todos los modelos debe dar la unidad, entonces

$$\sum_{j=1}^k P(M_j|D) = 1 = c_1 \sum_{j=1}^k e^{-0.5BIC_j}.$$

Note que

$$P(M_i|D) = \frac{e^{-0.5BIC_i}}{\sum_{j=1}^k e^{-0.5BIC_j}}$$

mediante la regla de Bayes.

La distribución de probabilidad de una nueva observación z será una mixtura o mezcla de distribuciones:

$$\sum P(M_i|D)f(z|M_i),$$

donde $f(z|M_i)$ es la densidad de la nueva observación de acuerdo al modelo M_i . Si la anterior distribución es la distribución que usará para hacer la predicción o pronóstico, entonces la esperanza de esa distribución será la usada para obtener la predicción puntual. Supongamos que estamos interesados en la predicción un paso adelante de la variable Z , es decir, $\hat{Z}_T(1)$. Sea $\hat{Z}_T^{(i)}(1)$ la predicción un paso adelante para el periodo $T+1$ con el modelo i , entonces $\hat{Z}_T(1)$ será:

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

Nota 7.52. Note que esta predicción es el resultado de ponderar las predicciones de todos los modelos por sus probabilidades para obtener una predicción agregada única. Esta forma de calcular las predicciones se conoce como promedio Bayesiano de modelo(BMA) por sus siglas en inglés. En general la predicción será mas precisa, en promedio, que la generada por los modelos individuales. Se puede cambiar la forma de ponderar, por ejemplo, si a cada modelo se le asigna la misma probabilidad $\frac{1}{k}$, entonces es el promedio simple la predicción. En general, es una combinación lineal de los pronósticos de cada modelo. Vale la pena decir, que la combinación podría no ser lineal en general, y usarse una estrategia de redes neuronales para obtener

la predicción. Finalmente, también se deben encontrar estrategias para construir intervalos de predicción para la predicción basada en la combinación de pronósticos.

8

Datos Atípicos, Datos Faltantes y Análisis Espectral de Series de Tiempo.

En éste capítulo, introduciremos unos temas importantes en el modelamiento de las series de tiempo: Datos atípicos u outliers, Datos faltantes y el análisis espectral de series de tiempo. Estos tópicos deberían ser revisados durante el análisis de una serie de tiempo debido a que pueden tener un gran impacto en el resultado final sino son incluidos.

8.1 Análisis de Intervención

Cómo se explica que [29] sección 12, las series de tiempo con bastante frecuencia se ven afectadas por sucesos puntuales conocidos. Por ejemplo, una pandemia, una huelga, un año bisiesto, un cambio en la ley, un accidente, o un cambio en festividad por nombrar algunos ejemplos. Al tenerse en cuenta estos sucesos en el modelamiento, se puede mejorar la precisión en la estimación de los parámetros y de los pronósticos.

Muchos de estos eventos pueden ser introducidas en el modelamiento a través de variables ficticias o variable de intervención. Las variables mas comunes son la variable impulso y escalón. La variable impulso representan fenómenos que ocurren únicamente en un instante, mientras que la variable escalón representan sucesos que comienzan en un instante conocido y se mantiene a través de ese instante.

Vamos a asumir que la serie $\{y_t\}$ está afectada en un instante dado, $t = h$ por un suceso conocido. Entonces en el tiempo h lo que observamos es la variable

$$z_h = \omega_0 + y_h$$

donde ω_0 es la magnitud del efecto sobre la serie. Si asumimos que la serie $\{y_t\}$ sigue un proceso ARIMA tal que se puede escribir como

$$y_t = \psi(B)a_t.$$

Para representar el instante de ocurrencia de este fenómeno, vamos a definir la variable impulso como la función indicadora

$$I_t^{(h)} = \begin{cases} 1, & \text{Si } t = h \\ 0, & \text{en otro caso.} \end{cases}$$

Con lo cual la serie observada $\{z_t\}$ sigue el modelo

$$z_t = \omega_0 I_t^{(h)} + y_t,$$

lo cual puede escribirse como:

$$z_t = \omega_0 I_t^{(h)} + \psi(B)a_t.$$

Para poder incluir efectos más complejos se requiere de introducir:

$$BI_t^{(h)} = I_{t-1}^{(h)} = I_t^{(h+1)}$$

y en forma general

$$B^j I_t^{(h)} = I_{t-j}^{(h)} = I_t^{(h+j)}.$$

Por qué?

Con esto, se pueden incluir intervenciones en m períodos subsecuentes como sigue:

$$z_t = \omega(B) I_t^{(h)} + \psi(B)a_t,$$

donde $\omega(B) = \omega_0 + \omega_1 B + \dots + \omega_m B^m$.

Si el número de períodos afectados es largo, la representación anterior requiere la estimación de muchos parámetros. Para esto, suele proponerse un modelo de pesos decrecientes como sigue:

$$z_t = \frac{\omega_0}{1 - \delta B} I_t^{(h)} + \psi(B)a_t,$$

con $0 < \delta < 1$, con lo cual $\frac{\omega_0}{1 - \delta B} = \omega_0(1 + \delta B + \delta^2 B^2 + \dots + \delta^m B^m + \dots)$. Ahora el efecto de la intervención es infinito pero tienden a disminuir a medida que pasa el tiempo. Este se conoce como cambio transitorio.

En ocasiones las intervenciones tienen un efecto que permanece en el tiempo. Estas intervenciones se modelan con la variable escalón, que se define como

$$S_t^{(h)} = \begin{cases} 0, & \text{Si } t < h \\ 1, & \text{Si } t \geq h. \end{cases}$$

en este caso el modelo queda como

$$z_t = \omega_0 S_t^{(h)} + \psi(B) a_t,$$

lo cual corresponde a un cambio de nivel, es decir, todo los valores posteriores al instante h están afectados por una cantidad constante ω_0 .

Finalmente, presentaremos un modelo para considerar un efecto gradual como el que sigue:

$$z_t = \sum_{j=0}^m \omega_j S_t^{(h+j)} + \psi(B) a_t,$$

en este caso los efectos son acumulativos y luego se estabilizan después de m períodos.

8.2 Datos Atípicos

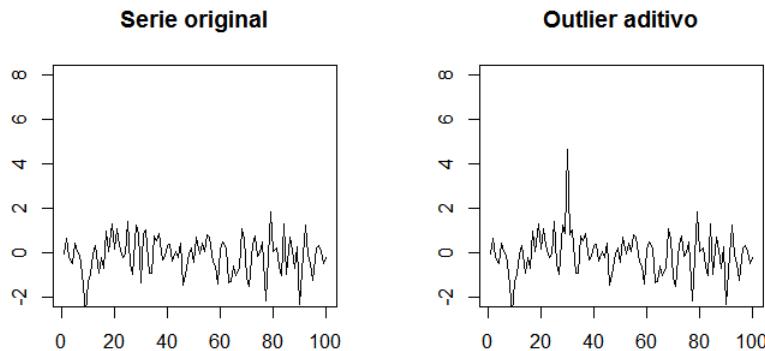
Definición 8.1. *Son valores que se alejan a más de tres desviaciones estándar de la media, los cuales no pueden ser expresados mediante un modelo. La presencia de estos datos se pueden deber a diferentes factores, por ejemplo la implementación de nueva maquinaria, errores en la medición o toma de datos, algún desastre natural o la implementación de una nueva política.*

Consecuencias de los Outliers

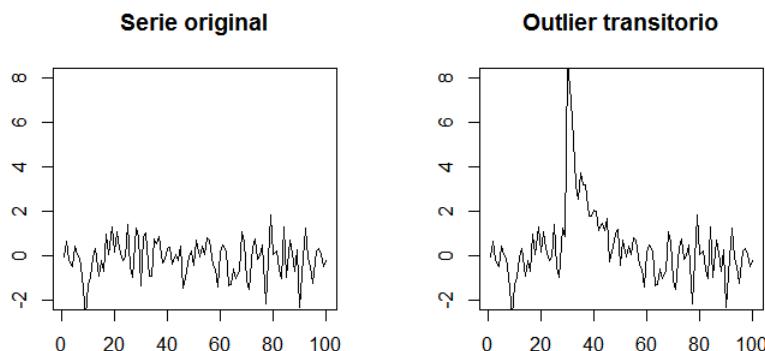
- ✓ Incrementan la varianza
- ✓ Pueden alterar la distribución original de los datos haciendo que no se cumplan ciertos supuestos
- ✓ Pueden sesgar o influenciar los resultados a un camino incorrecto,

Tipos de outliers

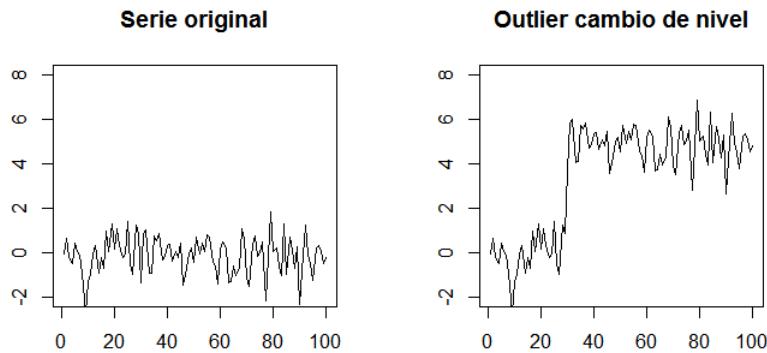
- ✓ **Outlier aditivo:** Se presenta cuando una observación tiene un valor muy grande o muy pequeño.



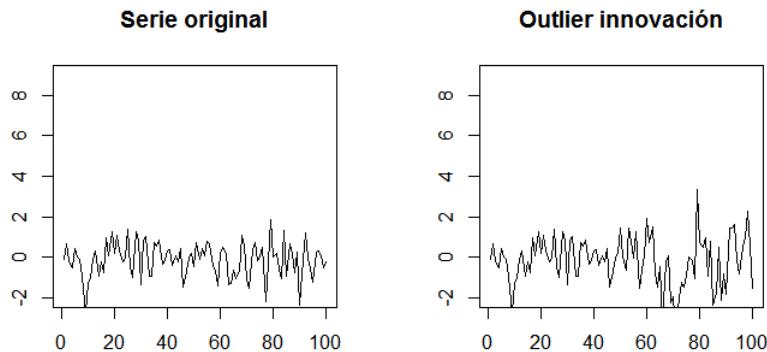
- ✓ **Outlier cambio de nivel:** Es un cambio permanente que afecta a todas las observaciones después del dato atípico.



- ✓ **Outlier cambio transitorio:** Semejante al cambio de nivel, afecta observaciones después del dato atípico, sin embargo en este caso su efecto disminuye rápidamente conforme las observaciones se alejan al impacto inicial.



- ✓ **Outlier innovación:** Es un cambio permanente en la serie, el cual puede ser descrito mediante un proceso.



En seguida veremos un algoritmo para la identificación de outliers en series de tiempo para modelos SARIMA, aunque la explicación será enfocada en modelos ARMA por facilidad.

Algoritmo de Chung, Chen y Lon-Mu Liu

Sea $\{Y_t\}$ una serie de tiempo que sigue un proceso ARMA(p,q), es decir:

$$Y_t = \frac{\theta(B)}{\alpha(B)\phi(B)} a_t$$

Donde $\theta(B)$, $\phi(B)$ y $\alpha(B)$ son polinomios de B, $a_t \approx RB(0, \sigma^2)$. Todas

las raíces de $\theta(B)$ y $\phi(B)$ están fuera del círculo unitario y las raíces de $\alpha(B)$ están dentro del círculo unitario.

Ahora suponga que Y_t tiene algún dato atípico, entonces se puede escribir como:

$$Y_t^* = Y_t + \omega \frac{A(B)}{G(B)H(B)} I_t^{(t)}.$$

ω es el impacto del dato atípico, y la función indicadora

$$I_t^{(h)} = \begin{cases} 1, & \text{Si } t = h \\ 0, & \text{en otro caso} \end{cases}$$

si la observación en el tiempo t es un dato atípico y $\frac{\theta(B)}{\alpha(B)\phi(B)}$ es la dinámica del outlier. Vale la pena decir que

$$BI_t^{(h)} = I_{t-1}^{(h)} = I_t^{(h+1)}$$

y en forma general

$$B^j I_t^{(h)} = I_{t-j}^{(h)} = I_t^{(h+j)}.$$

Observación: Dependiendo el tipo de dato atípico presente en la serie, se tendrá una dinámica diferente

Aditivo Transitorio Cambio de nivel Innovación

$$1 \quad \frac{1}{1 - \delta B} \quad \frac{1}{1 - B} \quad \frac{\theta(B)}{\phi(B)\alpha(B)}$$

Observación: Note que el outlier aditivo y el outlier cambio de nivel son un caso especial del outlier transitorio, en donde δ es 0 y uno respectivamente. En el Archivo Outliers.R se encuentra como llevar a cabo el análisis de outliers.

Ejercicio: Simule una un proceso ARMA(1,2) e introduzca un outlier aditivo de impacto 6.

Ejercicio: Simule una un proceso ARMA(2,1) e introduzca un outlier cambio de nivel de impacto 5.

Estimación y detección

Se define el polinomio $\pi(B) = \frac{\phi(B)\alpha(B)}{\theta(B)}$ para poder estimar los residuales contaminados por los outliers (\hat{e}_t) como $\pi(B)Y_t^*$, los cuales también pueden ser escritos como $\omega x_{it} + a_t$

Observación Como Y_t^* depende del tipo de outlier, \hat{e}_t también lo hará, con esta información es posible hallar la estimación para el impacto que tiene cada outlier, obteniendo:

$$\hat{\omega}_{AO}(t) = \frac{\sum_{t=t_1}^n \hat{e}_t x_{2t}}{\sum_{t=t_1}^n x_{2t}^2} \hat{\omega}_{LS}(t) = \frac{\sum_{t=t_1}^n \hat{e}_t x_{3t}}{\sum_{t=t_1}^n x_{3t}^2} \hat{\omega}_{TC}(t) = \frac{\sum_{t=t_1}^n \hat{e}_t x_{4t}}{\sum_{t=t_1}^n x_{4t}^2} \hat{\omega}_{IO}(t) = \hat{e}_{t_1}$$

En que: $x_{2(t_1+k)} = -\pi_k$, $x_{3(t_1+k)} = 1 - \sum_{j=1}^k \pi_j$ y $x_{4(t_1+k)} = \delta^k - \sum_{j=1}^{k-1} \delta^{k-j} \pi_j - \pi_k$

Una vez se estima el impacto de cada uno de los datos, se crea una estadística que ayude a tomar la decisión de cuales datos de la serie son atípicos, para esto se crean las siguientes estadísticas:

$$\begin{aligned}\hat{\tau}(t_1)_{AO} &= (\hat{\omega}_{AO}(t_1)/\hat{\sigma}_a) \left(\sum_{t=t_1}^n x_{2t}^2 \right)^{1/2} \\ \hat{\tau}(t_1)_{LS} &= (\hat{\omega}_{LS}(t_1)/\hat{\sigma}_a) \left(\sum_{t=t_1}^n x_{3t}^2 \right)^{1/2} \\ \hat{\tau}(t_1)_{TC} &= (\hat{\omega}_{TC}(t_1)/\hat{\sigma}_a) \left(\sum_{t=t_1}^n x_{4t}^2 \right)^{1/2} \\ \hat{\tau}(t_1)_{IO} &= \hat{\omega}_{IO}(t_1)/\hat{\sigma}_a\end{aligned}$$

Donde $\hat{\sigma}_a$ puede ser estimado mediante la desviación media absoluta, la $\alpha\%$ trimedia o el método omit-one.

- ✓ Una vez se ha decidido con que método estimar la varianza, se procede a calcular $\hat{\tau}_{AO}$, $\hat{\tau}_{LS}$, $\hat{\tau}_{TC}$ y $\hat{\tau}_{IO}$ para cada observación de la serie temporal
- ✓ Se toma el valor de $\hat{\tau}$ más alto para cada una de las observaciones, lo cual indicará cada observación que posible tipo de dato atípico es.
- ✓ Cada $\hat{\tau}$ se compara contra un valor critico (C), el cual será definido con anterioridad, si el valor de $\hat{\tau}(t) > C$, se dirá que la t -ésima observación es un dato atípico.

Observación: Por estudios de simulación se ha encontrado que un óptimo valor para C es 3.

8.3 Estimación o Predicción de Observaciones Faltantes

En el contexto de series de tiempo, los valores faltantes o ausentes deben ser manipulados de una manera adecuada. La idea más razonable consistirá en tratar de identificar un modelo para los datos en presencia de observaciones faltantes y luego usarlo para estimar los datos faltantes. Después de estimar los datos faltantes se procede a completar o imputar la serie con la estimación de los datos faltantes y se procede de nuevo a analizar la serie de tiempo sin la presencia de datos faltantes. Hay muchas metodologías para estimar o predecir los datos faltantes, sin embargo vamos a ver dos de éstas: vía el *filtro de Kalman* y usando *análisis de intervención*. Muchas otras metodologías son válidas también, sin embargo, las metodologías a tratar a acá son bastantes razonables. La filosofía de dato faltante acá considera que en un punto del tiempo, la variable aleatoria se realizó pero ésta no pudo ser registrada por algún motivo. Hay otras metodologías que consisten en que las ubicaciones de los datos faltantes son producto de un mecanismo probabilístico. Vamos a enfocar el problema de datos faltantes en modelos SARIMA.

8.3.1 Estimación de datos faltantes usando el Filtro de Kalman

Si hay datos faltantes en la serie, es decir si sólo observamos $Y_{i_1}, Y_{i_2}, \dots, Y_{i_r}$, $1 \leq i_1 < i_2 \dots < i_r \leq n$. Se introduce una nueva serie de tiempo $\{Y_t^*\}$, que tiene representación espacio estado como sigue:

$$Y_t^* = G_t^* X_t + W_t^*$$

donde

$$G_t^* = \begin{cases} G_t & \text{si } t \in \{i_1, \dots, i_r\}, \\ 0 & \text{en otro caso} \end{cases}$$

y

$$W_t^* = \begin{cases} W_t & \text{si } t \in \{i_1, \dots, i_r\}, \\ N_t & \text{en otro caso} \end{cases}$$

y la ecuación de estado

$$X_{t+1} = F_t X_t + V_t,$$

ver detalles en [2] páginas 483-488.

8.3.2 Estimación usando análisis de intervención

La idea de estimar un dato faltante usando análisis de intervención consiste en que en el tiempo t , donde hay un dato faltante, ese valor es llenado o imputado con un valor cualquiera (preferiblemente grande), y se procede a hacer una intervención en ese punto del tiempo como si hubiese un dato atípico. El efecto del intervención es estimado, y luego la estimación del dato faltante es el valor con que se llenó el dato faltante menos el efecto de la intervención.

8.4 Análisis Espectral de Series de Tiempo

El Espectro de un Proceso Estacionario

El periodograma para una frecuencia f_j es la medida de las varianzas debida al componente de esta frecuencia. Una forma alternativa de escribir el periodograma es:

$$I(f_j) = 2(\hat{\gamma}(0) + 2 \sum_{k=1}^{T-1} \hat{\gamma}(k) \cos(2\pi f_j k)) = 2\hat{\gamma}(0)(1 + 2 \sum_{k=1}^{T-1} \rho(k) \cos(2\pi f_j k)),$$

es decir, el periodograma es una combinación lineal de las autocovarianzas o de las autocorrelaciones muestrales. Esto nos lleva a pensar que el periodograma es una forma alternativa de representar la estructura de dependencia lineal observada en una serie.

Si hacemos $T \rightarrow \infty$, las propiedades de las series se aproximan a las del proceso estocástico que generó la serie y por lo tanto, las frecuencias básicas $\frac{1}{T} \leq f_j \leq \frac{1}{2}$ tenderán a cubrir todo el intervalo $0 \leq f \leq 1/2$, y el periodograma suavizado tenderá a una curva suave que dependerá del proceso generador temporal. Si el proceso que generó los datos es estacionario, las autocovarianzas muestrales $\hat{\gamma}(\cdot)$ tenderán a las teóricas del proceso estacionario $\gamma(\cdot)$, y así, se define el espectro del proceso estocástico mediante la expresión

$$S(f) = 2(\gamma(0) + 2 \sum_{k=1}^{\infty} \gamma(k) \cos(2\pi fk)),$$

para todo $0 \leq f \leq 1/2$, el cual existe para todo proceso estacionario y ergódico.

Tanto el espectro como la función de autocovarianza del proceso estocástico

recogen la misma información del proceso, es decir, si se conoce una de ellas se puede computarse la otra y al contrario. Los análisis basados en covarianzas se llaman análisis en el dominio del tiempo ya que las covarianzas definen la relación dinámica temporal entre las observaciones del proceso. En análisis basado en el espectro se denomina en el dominio de la frecuencia, ya que el espectro tiene en cuenta la contribución de los distintos armónicos a la variabilidad de la serie. La utilidad del espectro es que, como ocurre con el periodograma, podemos detectar los ciclos principales que forman el proceso. Note que la relación entre el periodograma y el espectro es similar a la del histograma y la función de densidad: la primera es una versión muestral, mientras que la segunda es la versión poblacional. Finalmente, se define la densidad espectral como

$$s(f) = \frac{S(f)}{\gamma(0)} = 2\left(1 + 2 \sum_{k=1}^{\infty} \rho(k) \cos(2\pi fk)\right).$$

El nombre de densidad es porque la función $s(f)$ tiene las mismas propiedades de una densidad de probabilidad. Esta función de densidad espectral representa la contribución relativa de los armónicos en cada intervalo a la variabilidad total del proceso.

Nota 8.2. Una forma de ver el espectro $S(f)$ es como la transformada de Fourier de la función de autocovarianza del proceso $\gamma(k)$, note que

$$S(f) = 2 \sum_{k=-\infty}^{\infty} \gamma(k) e^{-i2\pi fk},$$

ya que la exponencial compleja se puede escribir como $e^{-i2\pi fk} = \cos(2\pi fk) + i \operatorname{sen}(2\pi fk)$, ya que la suma de las funciones seno darán cero. Para que la propiedad se mantenga, de que varianza del proceso sea igual al área bajo la curva del espectro, ésta se define como

$$S(\omega) = \frac{1}{\pi} \sum_{k=-\infty}^{\infty} \gamma(k) e^{-i\omega k},$$

con $\omega = 2\pi f$.

en ocasiones se prefiere definir el espectro sobre el intervalo $(-\pi, \pi)$ en vez de $(0, \pi)$, y así su definición será

$$S^*(\omega) = \frac{1}{2\pi} \sum_{k=-\infty}^{\infty} \gamma(k) e^{-i\omega k}.$$

Adicionalmente puede demostrarse que a través de la transformada inversa de Fourier, podemos obtener la función de autocovarianzas del proceso con base el espectro

$$\gamma(k) = \int_{-\pi}^{\pi} e^{i\omega k} S(\omega) d\omega.$$

Finalmente, podemos verificar que cualquier proceso estocástico estacionario puede representarse como una suma infinita de funciones sinusoidales para todas las frecuencias en el intervalo $[0, 1/2]$ con amplitudes que son variables aleatorias independientes, es decir,

$$X_t = \int_{(-\pi, \pi]} e^{-i\omega k} dz(\omega),$$

donde $\{z(\omega), -\pi < \omega \leq \pi, \}$ es un proceso estocástico de valor complejo con incrementos ortogonales, la cual es conocida como la representación espectral del proceso $\{X_t\}$.

9

Modelo de Heterocedasticidad Condicional

En éste capítulo, introduciremos los primeros modelos no lineales de series de tiempo que son los modelos autoregresivos de heterocedasticidad condicional o modelos ARCH. .

9.1 Serie de Retornos

En muchas ocasiones, lo estudios financieros están basados en los retornos de las acciones y no en los precios directamente por básicamente dos razones:

- (Primera) Para el inversor, el retorno de una acción tiene el resumen completo de la oportunidad de inversión, adicional a que está libre de escala.
- (Segunda) La serie de retornos es más fácil de manejar que la serie de precio debido a que las propiedades estadísticas que poseen las series de retornos son mas atractivas.

Sin embargo, existen muchas definiciones de retornos.

Definiciones de retornos

Sea P_t el precio de una acción o de bienes básicos, un índice del mercado o la tasa de cambio. Daremos algunas definiciones de retornos que pueden ser útiles.

Retorno Simple en un periodo. Se define el retorno simple en un periodo como

$$R_t = \frac{P_t}{P_{t-1}} - 1 = \frac{P_t - P_{t-1}}{P_{t-1}}.$$

Retorno Simple en múltiples periodos. Se define el retorno simple para múltiples períodos como

$$1 + R_t[k] = \frac{P_t}{P_{t-k}} = \prod_{j=0}^{k-1} (1 + R_{t-j})$$

de acá se puede obtener lo que se llama el retorno mensualizado o anualizado.

Retorno compuesto continuamente. El logaritmo natural del retorno bruto simple de una acción es llamado el retorno compuesto continuamente o logaritmo del retorno

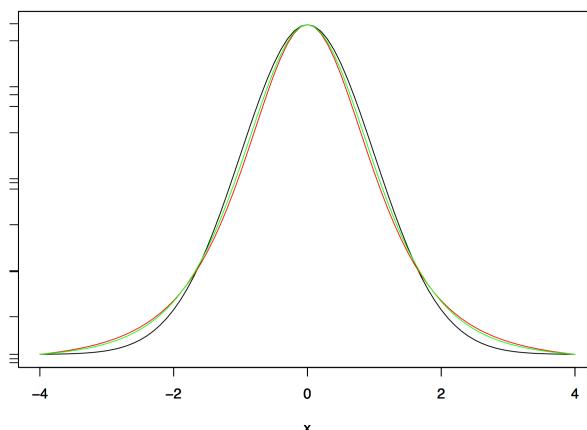
$$r_t = \ln(1 + R_t) = \ln \left(\frac{P_t}{P_{t-1}} \right) = p_t - p_{t-1}$$

donde $p_t = \ln(P_t)$.

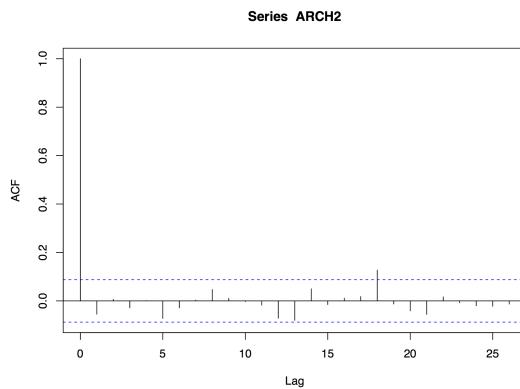
Propiedades Estadísticas de los Datos del Mercado Financiero.

Hechos Estilizados de las series financieras

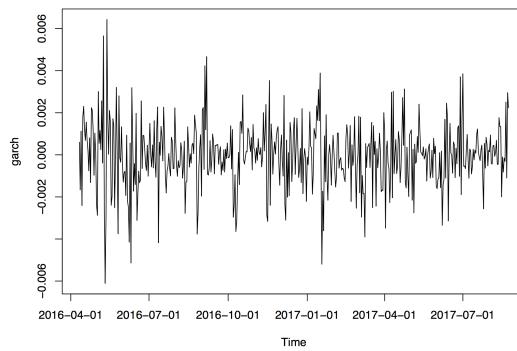
- (i) **Colas pesadas.** La distribución no condicional de los retornos tienen colas mas pesadas que la distribución normal.



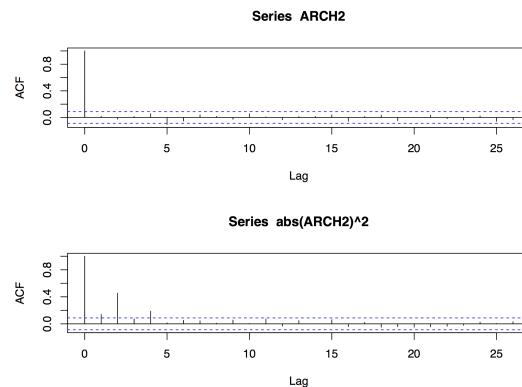
- (ii) **No autocorrelación serial.** Los retornos en general no presentan autocorrelación serial a menos que sean de alta frecuencia.



- (iii) **Asimetría.** La distribución no condicional es sesgada de forma negativa.
- (iv) **Clúster de datos extremos.** La volatilidad de los retornos es serialmente correlacionada, es decir, un retorno grande tiende a que se produzca un retorno grande también.

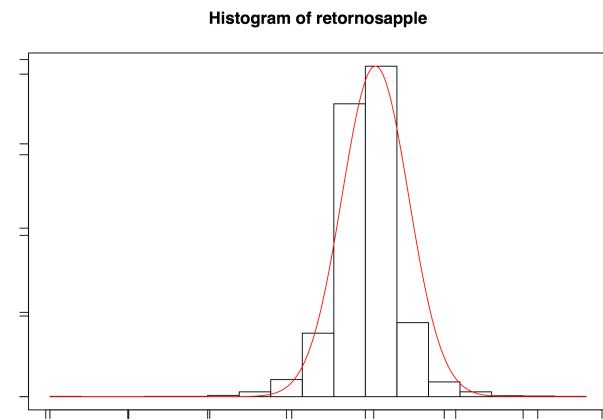
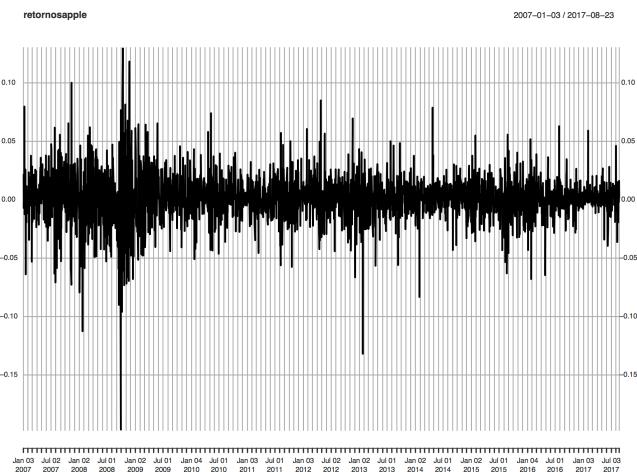


- (v) **Normalidad agregada.** Al aumentar la frecuencias de los retornos, se puede observar que la distribución tiende a parecerse mas a la distribución normal.
- (vi) **Efecto de Taylor.** Las autocorrelaciones para potencias de los retornos se muestran mas grandes que para los retornos para la potencia uno.

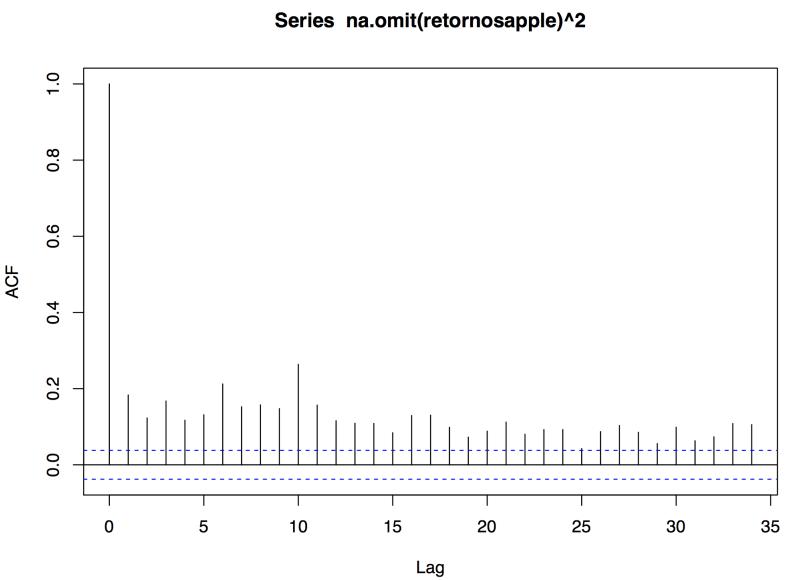
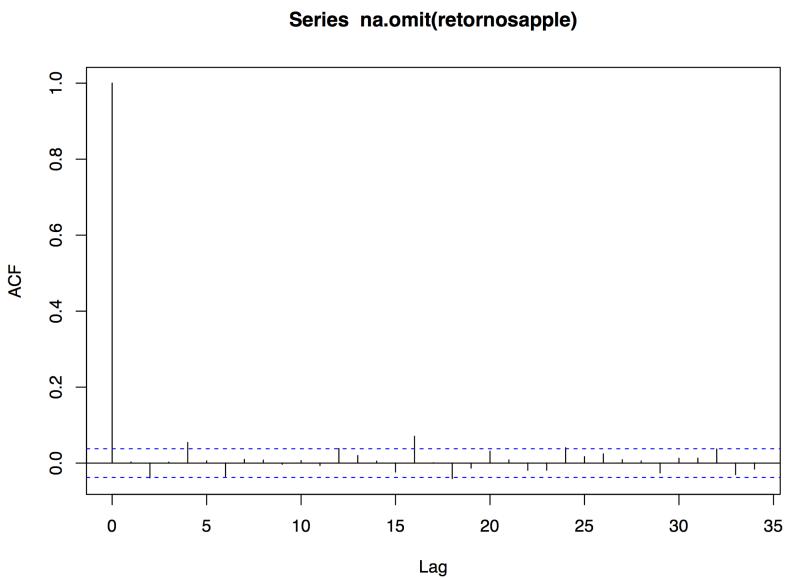


Ejemplo 9.1. Ejemplo Precio Acción de APPLE





La Kurtosis muestral para los retornos de la acción de apple es de 10.23795 y la asimetría de -0.4480229.



10

Series de tiempo Multivariadas

En este capítulo introduciremos los conceptos de procesos estocásticos multivariados y sus propiedades. Esto nos permitirá introducir algunos modelos de regresión de series de tiempo o regresión dinámica, en especial para series estacionarias. Para series no estacionarias es necesario involucrar el concepto de cointegración. Veamos unos ejemplos:

10.1 Preliminares

Un proceso estocástico vectorial K -dimensional o proceso estocástico multivariado es una función

$$\underline{Y} : \mathbb{Z} \times \Omega \rightarrow \mathbb{R}^K,$$

donde, para cada tiempo fijo $t \in \mathbb{Z}$, $\underline{Y}(t, \omega)$ es un vector aleatorio K -dimensional. Denotaremos \underline{Y}_t para indicar al vector aleatorio correspondiente al tiempo fijo $t \in \mathbb{Z}$, mientras que $\{\underline{Y}_t\}$ denotará al proceso estocástico.

Una realización de un proceso estocástico vectorial $\{\underline{Y}_t\}$ es una sucesión de vectores $\underline{Y}_t(\omega)$, para todo $t \in \mathbb{Z}$ y algún ω fijo. Es así como una serie de tiempo múltiple es considerada como una realización o posiblemente una parte finita de tal realización de un proceso estocástico, es decir, $\underline{Y}_1(\omega) = \underline{y}_1, \underline{Y}_2(\omega) = \underline{y}_2 \dots, \underline{Y}_T(\omega) = \underline{y}_T$. Al proceso estocástico subyacente de donde viene la realización se le llama *proceso generador de los datos*. El número de observaciones T es llamado tamaño de muestra o longitud de la serie de tiempo.

Supongamos que se desea obtener un pronóstico para el tiempo $T + h$ del vector de interés \underline{y} , cuyo periodo de tiempo final es T , ese pronóstico en

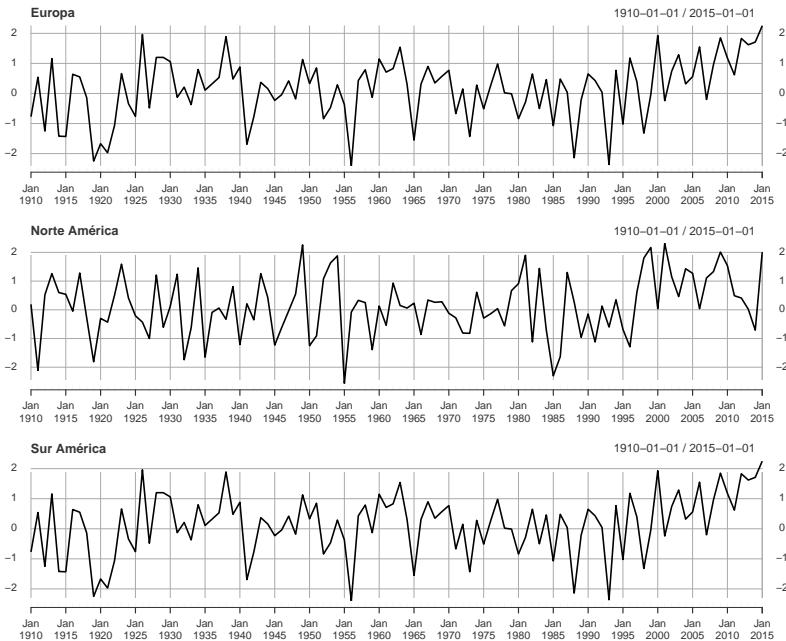


Figura 10.1: Temperatura anual promedio en diferentes sectores de la tierra

forma general puede tener la forma

$$\hat{y}_{T+h} = f(\underline{y}_T, \underline{y}_{T-1}, \dots)$$

para alguna función apropiada $f(\cdot)$, que depende de las observaciones pasadas $\underline{y}_T, \underline{y}_{T-1}, \dots$. Por supuesto uno de los objetivos de las series de tiempo consistirá en especificar formas “sensibles” de la función $f(\cdot)$. Por supuesto, tiene sentido empezar con funciones $f(\cdot)$ lineales, es decir, que los pronósticos dependen de forma lineal de las observaciones pasadas, mas específicamente

$$\hat{y}_{T+h} = \nu + A_1\underline{y}_T + A_2\underline{y}_{T-1} + \dots$$

Consideremos que sólo un número finito p de valores pasados son usados en la fórmula de la predicción y se está interesado en la predicción un paso adelante, es decir $h = 1$, entonces

$$\hat{y}_{T+1} = \nu + A_1\underline{y}_T + A_2\underline{y}_{T-1} + \dots + A_p\underline{y}_{T-p}.$$

Por supuesto el valor de pronóstico \hat{y}_{T+1} no es exactamente igual al valor \underline{y}_{T+1} , tal que denotamos al error de pronóstico(un paso adelante) como $u_{T+1} = \underline{y}_{T+1} - \hat{y}_{T+1}$. Con esto, es claro que

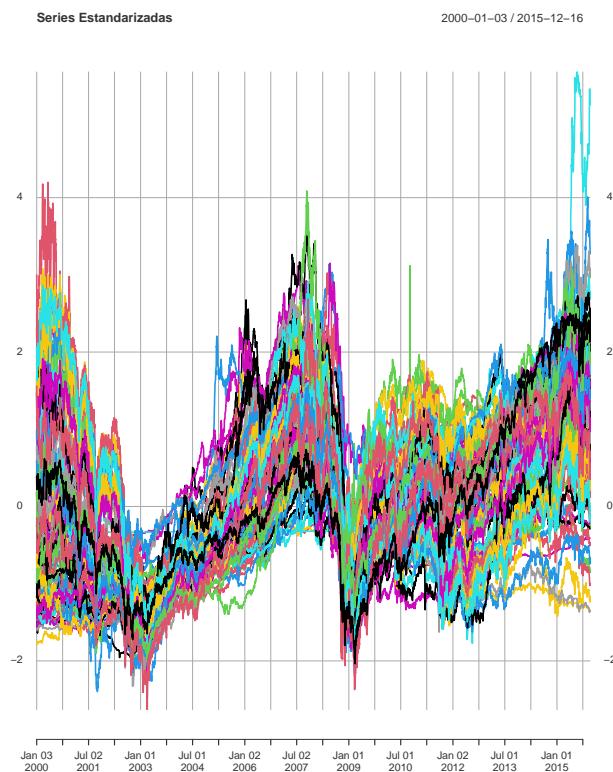


Figura 10.2: Precio de acciones estandarizadas

$$\underline{u}_{T+1} = \underline{y}_{T+1} - \hat{\underline{y}}_{T+1}$$

y así

$$\underline{y}_{T+1} = \hat{\underline{y}}_{T+1} + \underline{u}_{T+1} = \nu + A_1 \underline{y}_T + A_2 \underline{y}_{T-1} + \cdots + A_p \underline{y}_{T-p+1} + \underline{u}_{T+1} \quad (10.1)$$

Si asumimos que lo anterior son variables aleatorias, y que el mismo proceso generador de datos prevalece en cada periodo de tiempo T , entonces la ecuación (10.1) tiene la forma autorregresiva vectorial

$$\underline{Y}_t = \underline{\nu} + A_1 \underline{Y}_{t-1} + A_2 \underline{Y}_{t-2} + \cdots + A_p \underline{Y}_{t-p} + \underline{u}_t \quad (10.2)$$

donde las cantidades $\underline{Y}_t, \underline{Y}_{t-1}, \dots, \underline{Y}_{t-p}$ y \underline{u}_t son vectores aleatorios. Es razonable asumir que los errores de pronóstico \underline{u}_t son no autocorrelacionados, es decir, se asume que toda la información útil en el pasado de las \underline{Y}'_t s es usada en los pronósticos tal que no hay errores sistemáticos.

Si se especifica que $\underline{Y}_t = (\underline{Y}_{1,t}, \underline{Y}_{2,t}, \dots, \underline{Y}_{K,t})'$, $\underline{\nu} = (\nu_1, \nu_2, \dots, \nu_K)'$ y

$$A_i = \begin{bmatrix} a_{11,i} & a_{12,i} & \cdots & a_{1K,i} \\ a_{21,i} & a_{22,i} & \cdots & a_{2K,i} \\ \vdots & \vdots & \ddots & \vdots \\ a_{K1,i} & a_{K2,i} & \cdots & a_{KK,i} \end{bmatrix} \text{ y } \underline{u}_t = (\underline{u}_{1,t}, \dots, \underline{u}_{K,t})' \text{ entonces se puede}$$

ver que para cada componente $Y_{k,t}$, $k = 1, \dots, K$ del vector \underline{Y}_t , puede escribirse de la siguiente forma dada la ecuación autorregresiva (10.2):

$$\begin{aligned} Y_{k,t} = & \nu_k + a_{k1,1} Y_{1,t-1} + a_{k2,1} Y_{2,t-1} + \cdots + a_{kK,1} Y_{K,t-1} \\ & + a_{k1,2} Y_{1,t-2} + a_{k2,2} Y_{2,t-2} + \cdots + a_{kK,2} Y_{k,t-2} \\ & + \cdots + a_{k1,p} Y_{1,t-p} + a_{k2,p} Y_{2,t-p} + \cdots + a_{kK,p} Y_{k,t-p} + u_{k,t}. \end{aligned}$$

Se puede verificar que el predictor un paso adelante

$$\hat{\underline{Y}}_{T+1} = \nu + A_1 \underline{y}_T + A_2 \underline{y}_{T-1} + \cdots + A_p \underline{y}_{T-p+1}$$

se puede obtener del modelo autorregresivo en (10.2), el cual es un predictor óptimo en el sentido que tiene el menor error cuadrático medio de predicción, asumiendo que $\{\underline{u}_t\}$ es un proceso multivariado i.i.d. Esto es una primera forma de ver las generalizaciones que veremos en este curso, sin embargo dado que tenemos varias series de tiempo, debemos introducir unos primeros aspectos teóricos para poder estudiar estos modelos lineales de series de tiempo.

10.2 Procesos Estacionarios

Vamos a generalizar el concepto de estacionariedad al caso de procesos estocásticos multivariados. Para esto, consideremos el proceso estocástico

$$m\text{-dimensional } \{\underline{X}_t\}, \text{ tal que } \underline{X}_t = (X_{1,t}, \dots, X_{m,t})' = \begin{bmatrix} X_{1,t} \\ \vdots \\ X_{m,t} \end{bmatrix}.$$

Al igual que para el caso univariado, tenemos la función de medias vectorial

$$\underline{\mu}_t = E[\underline{X}_t] = \begin{bmatrix} E[X_{1,t}] \\ \vdots \\ E[X_{m,t}] \end{bmatrix} = \begin{bmatrix} \mu_{1,t} \\ \vdots \\ \mu_{m,t} \end{bmatrix}$$

y la función de autocovarianza matricial

$$\begin{aligned} \Gamma(t+h, t) &= Cov(\underline{X}_{t+h}, \underline{X}_t) = E[(\underline{X}_{t+h} - \underline{\mu}_{t+h})(\underline{X}_t - \underline{\mu}_t)'] \\ &= \begin{bmatrix} \gamma_{11}(t+h, t) & \gamma_{12}(t+h, t) & \cdots & \gamma_{1m}(t+h, t) \\ \vdots & \vdots & \ddots & \vdots \\ \gamma_{m1}(t+h, t) & \gamma_{m2}(t+h, t) & \cdots & \gamma_{mm}(t+h, t) \end{bmatrix} \end{aligned}$$

Vale la pena decir que el valor esperado de una matriz aleatoria consiste del valor esperado de cada una de sus componentes.

Definición 10.1. *El proceso estocástico m -dimensional $\{\underline{X}_t\}$ es estacionario en el sentido débil o débilmente estacionario si:*

- i) $\underline{\mu}_t$ no depende del tiempo t ,
- ii) $\Gamma(t+h, t)$ no depende del tiempo t para cada h .

Nota 10.2. *Si $\{\underline{X}_t\}$ es un proceso estacionario en el sentido débil entonces tenemos que:*

$$\underline{\mu} = E[\underline{X}_t] = \begin{bmatrix} \mu_1 \\ \vdots \\ \mu_m \end{bmatrix}$$

y

$$\Gamma(h) = E[(\underline{X}_{t+h} - \underline{\mu})(\underline{X}_t - \underline{\mu})'] = \begin{bmatrix} \gamma_{11}(h) & \gamma_{12}(h) & \cdots & \gamma_{1m}(h) \\ \vdots & \vdots & \ddots & \vdots \\ \gamma_{m1}(h) & \gamma_{m2}(h) & \cdots & \gamma_{mm}(h) \end{bmatrix}$$

en donde μ es la media del proceso estocástico y $\Gamma(h)$ es la matriz de autocovarianza en el rezago h . $\gamma_{ij}(h)$, para un rezago positivo h , se interpreta como una medida de dependencia lineal de la i -ésima componente $X_{i,t}$ y el h valor retardado de la j -ésima componente $X_{j,t-h}$.

Nota 10.3. Si el proceso $\{\underline{X}_t\}$ es estacionario en el sentido débil con función de autocovarianza matricial $\Gamma(\cdot)$, entonces para cada i , $\{X_{i,t}\}$ es un proceso estocástico estacionario univariado con función de autocovarianza $\gamma_{ii}(\cdot)$. La función $\gamma_{ij}(\cdot)$, $i \neq j$, se le conoce como función de autocovarianza cruzada de los procesos $\{X_{i,t}\}$ y $\{X_{j,t}\}$, donde $\gamma_{ij}(\cdot)$ y $\gamma_{ji}(\cdot)$ son en general diferentes. En efecto

$$\gamma_{ij}(h) = E[(X_{i,t+h} - \mu_i)(X_{j,t} - \mu_j)] (*)$$

mientras que

$$\gamma_{ji}(h) = E[(X_{j,t+h} - \mu_j)(X_{i,t} - \mu_i)] (**)$$

los cuales en general son diferentes puesto que en (*) la variable que está en el tiempo $t+h$ es X_i , mientras que en (**) la variable que está en el tiempo $t+h$ es X_j , lo cual hace que en general esas dos esperanzas no sean iguales.

A partir de la función de autocovarianza podemos definir la función de autocorrelación $R(\cdot)$ o $R_{\underline{X}}(\cdot)$ como sigue:

$$R(h) = \begin{bmatrix} \rho_{11}(h) & \rho_{12}(h) & \cdots & \rho_{1m}(h) \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{m1}(h) & \rho_{m2}(h) & \cdots & \rho_{mm}(h) \end{bmatrix}$$

donde $\rho_{ij}(h) = \gamma_{ij}(h)/[\gamma_{ii}(0)\gamma_{jj}(0)]^{1/2}$. Se puede verificar que

$$R(h) = D^{-1}\Gamma(h)D^{-1}$$

donde D es una matriz diagonal con las desviaciones estándar de las componentes de \underline{Y}_t , es decir, la diagonal de los documentos de D son las raíces cuadradas de los elementos de la diagonal de $\Gamma(0)$.

Nota 10.4. En efecto note que $D = \begin{bmatrix} \sqrt{\gamma_{11}(0)} & \cdots & 0 \\ \vdots & \ddots & \\ 0 & \cdots & \sqrt{\gamma_{mm}(0)} \end{bmatrix}$ y así

$$D^{-1} = \begin{bmatrix} 1/\sqrt{\gamma_{11}(0)} & \cdots & 0 \\ \vdots & \ddots & \\ 0 & \cdots & 1/\sqrt{\gamma_{mm}(0)} \end{bmatrix}.$$

Ahora nos enfocaremos en obtener la primera final de la matriz $R(h) = D^{-1}\Gamma(h)D^{-1}$. Es decir obtenemos

$$R(h) = \begin{bmatrix} 1/\sqrt{\gamma_{11}(0)} & \cdots & 0 \\ & \ddots & \\ 0 & \cdots & 1/\sqrt{\gamma_{mm}(0)} \end{bmatrix} \begin{bmatrix} \gamma_{11}(h) & \cdots & \gamma_{1m}(h) \\ \vdots & \ddots & \vdots \\ \gamma_{m1}(h) & \cdots & \gamma_{mm}(h) \end{bmatrix} \begin{bmatrix} 1/\sqrt{\gamma_{11}(0)} & \cdots & 0 \\ & \ddots & \\ 0 & \cdots & 1/\sqrt{\gamma_{mm}(0)} \end{bmatrix}$$

Para esto, note que la primera fila de esta matriz se obtiene de la siguiente manera:

$$\left[\frac{1}{\sqrt{\gamma_{11}(0)}} \gamma_{11}(h) \cdots \frac{1}{\sqrt{\gamma_{11}(0)}} \gamma_{1m}(h) \right] \begin{bmatrix} 1/\sqrt{\gamma_{11}(0)} & \cdots & 0 \\ & \ddots & \\ 0 & \cdots & 1/\sqrt{\gamma_{mm}(0)} \end{bmatrix}$$

lo cual es igual a

$$\left[\frac{\gamma_{11}(h)}{\sqrt{\gamma_{11}(0)\gamma_{11}(0)}} \cdots \frac{\gamma_{1m}(h)}{\sqrt{\gamma_{11}(0)\gamma_{mm}(0)}} \right] = [\rho_{11}(h) \cdots \rho_{1m}(h)]$$

la cual coincide con la definición de $\rho_{ij}(h)$ dada anteriormente.

Ejemplo 10.5. Considere el proceso estacionario bivariado $\{\underline{X}_t\}$ definido por

$$X_{1,t} = Zt$$

$$X_{2,t} = Z_t + 0.75Z_{t-10}$$

donde $\{Z_t\} \sim RB(0, 1)$.

Computemos el valor esperado del proceso $\{\underline{X}_t = (X_{1,t}, X_{2,t})'\}$.

$$E[\underline{X}_t] = (E[X_{1,t}], E[X_{2,t}])' = (E[Zt], E[Z_t + 0.75Z_{t-10}])',$$

usando el hecho que el proceso $\{Z_t\}$ es ruido blanco, tenemos que

$$E[\underline{X}_t] = [0, 0]' = \underline{\mu}.$$

Ahora calculemos $E[(\underline{X}_{t+h} - \underline{\mu})(\underline{X}_t - \underline{\mu})'] = E[\underline{X}_{t+h}\underline{X}_t']$. Vemos ahora que

$$E \left[\begin{pmatrix} Z_{t+h} \\ Z_{t+h} + 0.75Z_{t+h-10} \end{pmatrix} \begin{pmatrix} Z_t & Z_t + 0.75Z_{t-10} \end{pmatrix} \right]$$

con la cual, haciendo operaciones tenemos

$$E \left[\begin{pmatrix} Z_{t+h}Z_t & Z_{t+h}(Z_t + 0.75Z_{t-10}) \\ (Z_{t+h} + 0.75Z_{t+h-10})Z_t & (Z_{t+h} + 0.75Z_{t+h-10})(Z_t + 0.75Z_{t-10}) \end{pmatrix} \right].$$

Note que para $h = 0$, tenemos entonces $\gamma(0) = \begin{bmatrix} 1 & 1 \\ 1 & 1.5625 \end{bmatrix}$.

Para $h = 10$, tenemos entonces $\gamma(10) = \begin{bmatrix} 0 & 0 \\ 0.75 & 0.75 \end{bmatrix}$.

Para $h = -10$, tenemos entonces $\gamma(-10) = \begin{bmatrix} 0 & 0.75 \\ 0 & 0.75 \end{bmatrix}$.

Se puede verificar que para otro rezago $h \neq 0, 10$, $\Gamma(h) = 0$. Note que las función de autocorrelación matricial se puede construir basado en la matriz D^{-1} , cual es obtenida a partir de las desviaciones estándar de la matriz $\Gamma(0)$, así que $D^{-1} = \begin{bmatrix} 1 & 0 \\ 0 & \frac{1}{\sqrt{1.5625}} \end{bmatrix}$. Con esto podemos calcular $R(h) = D^{-1}\Gamma(h)D^{-1}$ para cada h , y quedan:

$$R(0) = \begin{bmatrix} 1 & 0.8 \\ 0.8 & 1 \end{bmatrix}, R(10) = \begin{bmatrix} 1 & 0 \\ 0.6 & 0.48 \end{bmatrix} \text{ y } R(-10) = \begin{bmatrix} 1 & 0.6 \\ 0 & 0.48 \end{bmatrix}.$$

Note que en el ejemplo anterior se puede evidenciar que $\Gamma(h) = \Gamma(-h)'$.

Vamos ahora a definir el proceso ruido blanco multivariado K -dimensional.

Proposición 10.6. *La función de autocovarianza matricial $\Gamma(h)$ de un proceso estocástico estacionario K -dimensional \underline{X}_t tiene las siguientes propiedades*

i) $\Gamma(h) = \Gamma(-h)'$,

ii) $|\gamma_{ij}(h)| \leq [\gamma_{ii}(0)\gamma_{jj}(0)]^{1/2}$, $i, j = 1, \dots, K$,

iii) $\gamma_{ii}(h)$ es una función de autocovarianza, $i = 1, \dots, K$, y

iv) $\sum_{j,k=1}^n \mathbf{a}'_j \Gamma(j-k) \mathbf{a}_k \geq 0$, para todo $n \in \{1, 2, \dots\}$, $\mathbf{a}_1, \dots, \mathbf{a}_n \in \mathbb{R}^K$.

Nota 10.7. Note que la proposición anterior es una generalización de las propiedades de la función de autocovarianza al caso multivariado, sin embargo acá queremos detallar la primera propiedad:

$$\begin{aligned} \Gamma(h) &= E[(\underline{X}_{t+h} - \underline{\mu})(\underline{X}_t - \underline{\mu})'] \\ &= E[(\underline{X}_t - \underline{\mu})(\underline{X}_{t-h} - \underline{\mu})'] \quad \text{Por la estacionariedad} \\ &= E\left[\left\{(\underline{X}_{t-h} - \underline{\mu})(\underline{X}_t - \underline{\mu})'\right\}'\right] \quad \text{Por } \mathbf{C} = (\mathbf{C}')' \\ &= [\Gamma(-h)]' \\ &= \Gamma(-h). \end{aligned}$$

Definición 10.8. *Un proceso estocástico vectorial o multivariado K -dimensional $\{\underline{u}_t\}$ se dice ruido blanco o proceso de innovaciones, lo cual escribiremos $\{\underline{u}\} \sim RB(\underline{0}, \Sigma_{\underline{u}})$, si*

- i) $E[\underline{u}_t] = \underline{Q}$ para cada tiempo t .
- ii) $E[\underline{u}_t \underline{u}'_t] = \Sigma_u$ y $E[\underline{u}_t \underline{u}'_s] = \mathbf{0}$ para $s \neq t$

Tanto Σ_u como \underline{Q} son matrices de orden $K \times K$, y se pide que Σ_u sea no singular, es decir que tenga inversa o que sea definida positiva.

Ahora introduciremos la versión multivariada de proceso i.i.d.

Definición 10.9. Un proceso estocástico m -dimensional $\{\underline{Z}_t\}$ es llamado ruido i.i.d con media \underline{Q} y matriz de covarianza $\Sigma_{\underline{Z}}$, lo cual escribiremos como $\{\underline{Z}_t\} \sim i.i.d(\underline{Q}, \Sigma_{\underline{Z}})$, si los vectores aleatorios $\{\underline{Z}_t\}$ son independientes e idénticamente distribuidos(es decir, cada conjunto finito de vectores aleatorios formado por vectores del proceso son i.i.d) con media \underline{Q} y matriz de covarianza $\Sigma_{\underline{Z}}$.

Ahora veremos la generalización del proceso lineal al caso multivariado. Es mas frecuente encontrar que la estructura de las series de tiempo multivariadas sean no-lineales, sin embargo los modelos lineales puede entregar aproximaciones adecuadas para inferencia y pronóstico.

Definición 10.10. Un proceso estocástico K -dimensional $\{\underline{X}_t\}$ es un proceso lineal si tiene representación

$$\underline{X}_t = \sum_{j=-\infty}^{\infty} \mathbf{C}_j \underline{Z}_{t-j}, \quad \{\underline{Z}_t\} \sim RB(\underline{Q}, \Sigma_{\underline{Z}}) \quad (10.3)$$

donde \mathbf{C}_j es una sucesión de matrices $K \times K$ cuyas componentes son absolutamente sumables y la matriz $\mathbf{C}_0 = I_K$. También se puede considerar que las matrices de coeficientes deben satisfacer que:

$$\sum_{j=-\infty}^{\infty} \|\mathbf{C}_j\| < \infty,$$

donde $\|\mathbf{A}\|$ denota la norma de la matriz \mathbf{A} , la cual por ejemplo puede ser la norma de Frobenius $\|\mathbf{A}\| = \sqrt{\text{tr}(\mathbf{A}\mathbf{A}')}$. Ver libro [27] pág. 687, para mostrar que las dos definiciones son equivalentes.

Note el proceso lineal está bien definido en virtud de la proposición C.9 de la Página 688 del libro [27], la cual consiste de la existencia de la suma infinita de variables aleatorias ver 10.3. Adicionalmente podemos encontrar la forma de computar los momentos del proceso lineal basado en la proposición C.10 ver 10.4 , la cual es una forma general de obtener los momentos de sumas infinitas de vectores aleatorios.

Proposition C.9 (*Existence of Infinite Sums of Random Vectors*)

Suppose $\{A_i\}$ is an absolutely summable sequence of real $(K \times K)$ matrices and $\{z_t\}$ is a sequence of K -dimensional random variables satisfying

$$E(z'_t z_t) \leq c, \quad t = 0, \pm 1, \pm 2, \dots,$$

for some finite constant c . Then there exists a sequence of K -dimensional random variables $\{y_t\}$ such that

$$\sum_{i=-n}^n A_i z_{t-i} \xrightarrow[n \rightarrow \infty]{q.m.} y_t.$$

The sequence is uniquely determined except on a set of probability zero. ■

Figura 10.3:

Proposition C.10 (*Moments of Infinite Sums of Random Vectors*)

Suppose z_t satisfies the conditions of Proposition C.9, $\{A_i\}$ and $\{B_i\}$ are absolutely summable sequences of $(K \times K)$ matrices,

$$y_t = \sum_{i=-\infty}^{\infty} A_i z_{t-i} \quad \text{and} \quad x_t = \sum_{i=-\infty}^{\infty} B_i z_{t-i}.$$

Then

$$E(y_t) = \lim_{n \rightarrow \infty} \sum_{i=-n}^n A_i E(z_{t-i})$$

and

$$E(y_t x'_t) = \lim_{n \rightarrow \infty} \sum_{i=-n}^n \sum_{j=-n}^n A_i E(z_{t-i} z'_{t-j}) B'_j,$$

where the limit of the sequence of matrices is the matrix of limits of the sequences of individual elements. ■

Figura 10.4:

Con las anteriores proposiciones podemos obtener los momentos del proceso lineal.

Proposición 10.11. *La media del proceso lineal definido en 10.3 el vector \underline{Q} mientras que la función de autocovarianza matricial es*

$$\Gamma(h) = \text{Cov}(\underline{X}_{t+h}, \underline{X}_t) = \sum_{j=-\infty}^{\infty} \mathbf{C}_{j+h} \Sigma_{\mathcal{Z}} \mathbf{C}_j'.$$

Nota 10.12. *Es claro que el proceso lineal definido anteriormente tiene vector medias cero, sin embargo esta definición puede extenderse para considerar una media diferente del vector cero, es decir, considerar un vector de medias $\underline{\mu} \neq \underline{Q}$ como sigue:*

$$\underline{X}_t = \underline{\mu} + \sum_{j=-\infty}^{\infty} \mathbf{C}_j \mathcal{Z}_{t-j}, \quad \{\mathcal{Z}_t\} \sim RB(\underline{Q}, \Sigma_{\mathcal{Z}}). \quad (10.4)$$

En este caso $E[\underline{X}_t] = \underline{\mu}$, mientras que la FACV matricial no se altera y queda como en la proposición 10.11.

10.2.1 Estimación del vector de medias y la FACV matricial

Data una muestra $\underline{X}_1, \dots, \underline{X}_n$ de un proceso estocástico estacionario, consideremos los estimadores del vector de medias $\underline{\mu}$ y de la función de autocovarianza matricial $\Gamma(h)$ como

$$\hat{\underline{\mu}} = \bar{\underline{X}} = \frac{1}{n} \sum_{t=1}^n \underline{X}_t$$

y

$$\hat{\Gamma}(h) = \frac{1}{n} \sum_{t=1}^{n-h} (\underline{X}_{t+h} - \bar{\underline{X}})(\underline{X}_t - \bar{\underline{X}})'.$$

De aquí se puede verificar que el estimador de la función de autocorrelación cruzada es

$$\hat{\rho}_{ij}(h) = \frac{\hat{\gamma}_{ij}(h)}{(\hat{\gamma}_{ii}(0)\hat{\gamma}_{jj}(0))^{1/2}}.$$

En el caso que $i = j$ tenemos $\hat{\rho}_{ii}(h)$ que es la FAC muestral de la i -ésima serie, con lo cual tenemos que un estimador de la matriz de autocorrelaciones está dado por

$$\hat{R}(h) = \hat{D}^{-1} \hat{\Gamma}(h) \hat{D}^{-1}.$$

Proposition 8.3.1. *If $\{\mathbf{X}_t\}$ is a stationary multivariate time series with mean $\boldsymbol{\mu}$ and covariance function $\Gamma(\cdot)$, then as $n \rightarrow \infty$,*

$$E(\bar{\mathbf{X}}_n - \boldsymbol{\mu})'(\bar{\mathbf{X}}_n - \boldsymbol{\mu}) \rightarrow 0 \quad \text{if } \gamma_{ii}(n) \rightarrow 0, \quad 1 \leq i \leq m,$$

and

$$nE(\bar{\mathbf{X}}_n - \boldsymbol{\mu})'(\bar{\mathbf{X}}_n - \boldsymbol{\mu}) \rightarrow \sum_{i=1}^m \sum_{h=-\infty}^{\infty} \gamma_{ii}(h) \quad \text{if } \sum_{h=-\infty}^{\infty} |\gamma_{ii}(h)| < \infty, \quad 1 \leq i \leq m.$$

Figura 10.5:

La proposición siguiente 10.5 nos entrega algunos resultados asintóticos de la media muestral, la cual es sacada del libro [3] Página 237.

Si se imponen supuestos mas fuertes sobre el proceso $\{\underline{X}_t\}$, se puede mostrar que la distribución de $\bar{\underline{X}}$ es aproximadamente normal multivariada para un tamaño de muestra grande.

En las misma página del libro se puede encontrar la forma de construir intervalos de confianza individuales y regiones de confianza para la media del proceso.

Basados en la proposición 10.6 de la página 238 del libro [3], tenemos la distribución de la FAC cruzada muestral, bajo el supuesto que las dos series consideradas son independientes. Si se consideran que las series tienen alguna estructura de correlación, entonces se debe aplicar un resultado análogo a la fórmula de Bartlett del caso univariado que se encuentra en la Pág.242 del libro [3] para el caso de la FAC cruzada(FACC) y el cual referenciamos en 10.7.

Nota 10.13. *De la proposición se puede crear un procedimiento para probar la correlación lineal entre dos series de tiempo. Si una de las dos series es ruido blanco, entonces se puede verificar que $\hat{\rho}_{12}(h)$ es aproximadamente normal con media 0 y varianza $1/n$, en tal caso es directo probar la hipótesis que $\rho_{12}(h) = 0$, como se hace en el caso de la autocorrelación de una serie univariada. Sin embargo, si ninguno de los procesos es ruido blanco, entonces es posible que un valor $\hat{\rho}_{12}(h)$ que es grande, relativo a $n^{-1/2}$, no necesariamente indica que $\rho_{12}(h)$ es diferente de cero, lo cual puede llevar a relaciones espurias. Ver comentario Página 239 libro [3].*

Nota 10.14. Prueba de independencia de dos series de tiempo estacionarias Dado que la distribución asintótica de $\hat{\rho}_{ij}(h)$ depende de $\rho_{11}(\cdot)$ y $\rho_{22}(\cdot)$, cualquier prueba de independencia de dos series de tiempo no se puede basar únicamente en los valores estimados de la FAC cruzada $\rho_{12}(h)$, sin

Theorem 8.3.1. Let $\{\mathbf{X}_t\}$ be the bivariate time series whose components are defined by

$$X_{t1} = \sum_{k=-\infty}^{\infty} \alpha_k Z_{t-k,1}, \quad \{Z_{t1}\} \sim \text{IID}(0, \sigma_1^2),$$

and

$$X_{t2} = \sum_{k=-\infty}^{\infty} \beta_k Z_{t-k,2}, \quad \{Z_{t2}\} \sim \text{IID}(0, \sigma_2^2),$$

where the two sequences $\{Z_{t1}\}$ and $\{Z_{t2}\}$ are independent, $\sum_k |\alpha_k| < \infty$, and $\sum_k |\beta_k| < \infty$.

Then for all integers h and k with $h \neq k$, the random variables $n^{1/2} \hat{\rho}_{12}(h)$ and $n^{1/2} \hat{\rho}_{12}(k)$ are approximately bivariate normal with mean $\mathbf{0}$, variance $\sum_{j=-\infty}^{\infty} \rho_{11}(j) \rho_{22}(j)$, and covariance $\sum_{j=-\infty}^{\infty} \rho_{11}(j) \rho_{22}(j+k-h)$, for n large.

[For a related result that does not require the independence of the two series $\{X_{t1}\}$ and $\{X_{t2}\}$ see Bartlett's Formula, Section 8.3.4 below.]

Figura 10.6:

Bartlett's Formula:

If $\{\mathbf{X}_t\}$ is a bivariate Gaussian time series with covariances satisfying $\sum_{h=-\infty}^{\infty} |\gamma_{ij}(h)| < \infty$, $i, j = 1, 2$, then

$$\lim_{n \rightarrow \infty} n \text{Cov}(\hat{\rho}_{12}(h), \hat{\rho}_{12}(k)) = \sum_{j=-\infty}^{\infty} \left[\rho_{11}(j) \rho_{22}(j+k-h) + \rho_{12}(j+k) \rho_{21}(j-h) - \rho_{12}(h) \{ \rho_{11}(j) \rho_{12}(j+k) + \rho_{22}(j) \rho_{21}(j-k) \} - \rho_{12}(k) \{ \rho_{11}(j) \rho_{12}(j+h) + \rho_{22}(j) \rho_{21}(j-h) \} + \rho_{12}(h) \rho_{12}(k) \left\{ \frac{1}{2} \rho_{11}^2(j) + \rho_{12}^2(j) + \frac{1}{2} \rho_{22}^2(j) \right\} \right]$$

Corollary 8.3.1. If $\{\mathbf{X}_t\}$ satisfies the conditions for Bartlett's formula, if either $\{X_{t1}\}$ or $\{X_{t2}\}$ is white noise, and if

$$\rho_{12}(h) = 0, \quad h \notin [a, b],$$

then

$$\lim_{n \rightarrow \infty} n \text{Var}(\hat{\rho}_{12}(h)) = 1, \quad h \notin [a, b].$$

Figura 10.7:

tener en cuenta la naturaleza de las dos series. Entonces un procedimiento que permita probar la independencia de dos series de tiempo requiere de un pre-blanqueamiento de las dos series antes de computar la FACC $\hat{\rho}_{12}(h)$. Este pre-blanqueamiento se puede llevar a cabo de la siguiente manera (El pre-blanqueamiento en su forma mas general consiste en la aplicación de un filtro lineal):

- 1) Ajustar modelos ARMA a cada una de las series $\{X_{it}\}$, para $i = 1, 2$.
- 2) Obtener los residuales $\{\hat{W}_{i,t}\}$ para $i = 1, 2$ basados en los modelos ajustados del paso 1).
- 3) Si los modelos ARMA ajustados son los correctos, entonces los residuales $\{\hat{W}_{i,t}\}$ para $i = 1, 2$ deberían ser una sucesión de ruidos blanco.

Ahora para probar la hipótesis H_0 que $\{X_{1t}\}$ y $\{X_{2t}\}$ son series independientes, entonces bajo H_0 , las series $\{Z_{1t}\}$ y $\{Z_{2t}\}$ son también independientes. Así que en virtud del resultado 10.6 tenemos que las FACC muestrales $\hat{\rho}_{12}(h)$ y $\hat{\rho}_{12}(k)$, para $h \neq k$ de $\{Z_{1t}\}$ y $\{Z_{2t}\}$ son para n lo suficientemente grande, independientes y normalmente distribuidas con medias 0 y varianzas $1/n$. Con esto presente, tenemos que una prueba de independencia aproximada consistirá en comparar los valores $|\hat{\rho}_{12}(h)|$ basados en los residuales en 2) con $1.96n^{1/2}$, tal como se hace para chequear no autocorrelación de una serie univariada. También se puede hacer un pre-blanqueo de una sola serie, y continuar con el mismo procedimiento de prueba ver Página 239 libro [3].

Cuando la dimensión es grande, es difícil poder chequear simultáneamente y comprender las correlaciones cruzadas. Para esto se utiliza la matriz simplificada propuesta por **Tiao y Box**. Para cada FAC matricial muestral $\hat{\Gamma}(h)$, se define la matriz simplificada $s(h) = [s_{ij}(h)]$ como:

$$s_{ij}(h) = \begin{cases} + & \text{si } \hat{\rho}_{ih}(h) \geq 1.96/\sqrt(n) \\ - & \text{si } \hat{\rho}_{ih}(h) \leq -1.96/\sqrt(n) \\ \cdot & \text{si } |\hat{\rho}_{ih}(h)| < 1.96/\sqrt(n) \end{cases}$$

La cual provee un resumen de la FACC muestral en cada rezago a un nivel de significancia aproximado del 5%. En seguida veremos una prueba global de correlación cruzada cero.

10.2.2 Prueba Portmanteau para Correlación Cruzada Cero

Inicialmente, es interesante chequear la existencia de una dinámica de dependencia lineal en los datos. Es decir, se debe verificar la hipótesis nula $H_0 : R(1) = R(2) = \dots = R(l) = \mathbf{0}$ versus la hipótesis alternativa $H_a : R(i) \neq \mathbf{0}$ para algún rezago $0 \leq i \leq l$, para algún entero positivo l . Hay varias pruebas, pero expondremos la prueba de Ljung-Box multivariada, la cual se basa en el estadístico de prueba

$$Q_k(l) = T^2 \sum_{j=1}^l \frac{1}{T-j} \text{tr} \left(\hat{\Gamma}'(j) \hat{\Gamma}'(0)^{-1} \hat{\Gamma}(j) \hat{\Gamma}(0)^{-1} \right) \quad (10.5)$$

donde k es la dimensión del vector de las series y l es el número de rezagos incluidos en la prueba. El estadístico puede ser escrito equivalentemente con base en la matriz de autocorrelaciones como sigue:

$$Q_k(l) = T^2 \sum_{j=1}^l \frac{1}{T-j} \hat{\mathbf{b}}'(j) (\hat{R}^{-1}(0) \otimes \hat{R}^{-1}(0)) \hat{\mathbf{b}}(j)$$

Bajo la hipótesis nula, y con la condición que el proceso es Gaussiano, entonces la estadística $Q_k(l)$ tiene distribución asintótica $\chi_{lk^2}^2$.

10.2.3 Pronóstico

Uno de los objetivos del análisis de series de tiempo es la predicción. Supongamos que se está interesado en predecir el valor Z_{T+h} basados en la información disponible hasta el tiempo T . Tal predicción se conoce como el pronóstico h pasos adelante de la serie en el índice del tiempo T . T es el origen del pronóstico, mientras que h es el horizonte de pronóstico. Notemos a F_t la información disponible hasta el tiempo t , la cual consiste de las observaciones Z_1, \dots, Z_t . En la práctica, el procesos generador de datos es desconocido, y se debe usar la información F_T que permita construir un modelo estadístico para la predicción. Se debe asumir que el modelo usado en predicción es el verdadero modelo generador de los datos. Con esto, se debe considerar que los pronósticos producidos por cualquier método que asume que el modelo ajustado es el verdadero modelo, van probablemente a sub-estimar la verdadera variabilidad de la serie de tiempo.

Por supuesto que los pronósticos también depende de la función de pérdida. Recordemos que la función de pérdida se define de la siguiente manera siguiendo a [12] y la teoría de la decisión:

- a) Un dominio de observación, \mathbf{O} , el cual compromete los resultados potenciales de una característica.
- b) Una clase \mathcal{F} de medidas de probabilidad sobre el dominio de observación, la cual constituye una familia de distribuciones de probabilidad para la observación futura.
- c) Un dominio de acciones, \mathbf{A} , el cual compromete las acciones potenciales del que toma las decisiones.
- d) Una función de pérdida $L : \mathbf{A} \times \mathbf{O} \rightarrow [0, +\infty)$, donde $L(a, o)$ representa la pérdida incurrida cuando el que toma las decisiones toma la acción $a \in \mathbf{A}$, y la observación $o \in \mathbf{O}$ se materializa.

Sin embargo siguiendo a [37], se propone la minimización del error cuadrático medio(MSE) de predicción, el cual consiste en lo siguiente: sea $\hat{\mathcal{Z}}_{T+h}$ un pronóstico arbitrario de Z_{T+h} con origen de pronóstico en T . Entonces el error de pronóstico es dado como $Z_{T+h} - \hat{\mathcal{Z}}_{T+h}$, y el error cuadrático medio del error de pronóstico es dado por

$$MSE(\hat{\mathcal{Z}}_{T+h}) = E \left[(Z_{T+h} - \hat{\mathcal{Z}}_{T+h}) (Z_{T+h} - \hat{\mathcal{Z}}_{T+h})' \right],$$

en donde el $MSE(\hat{\mathcal{Z}}_{T+h})$ es una matriz. No es difícil verificar que si consideramos la esperanza condicional $E [Z_{T+h}|F_t]$, entonces se puede verificar ver Página 16 libro [37] que

$$MSE(\hat{\mathcal{Z}}_{T+h}) = MSE(E [Z_{T+h}|F_t]) + E \left[(E [Z_{T+h}|F_t] - \hat{\mathcal{Z}}_{T+h}) (E [Z_{T+h}|F_t] - \hat{\mathcal{Z}}_{T+h})' \right].$$

Ahora, dado que se puede verificar que $E \left[(E [Z_{T+h}|F_t] - \hat{\mathcal{Z}}_{T+h}) (E [Z_{T+h}|F_t] - \hat{\mathcal{Z}}_{T+h})' \right]$ es definida no negativa, entonces se puede concluir que

$$MSE(\hat{\mathcal{Z}}_{T+h}) \geq MSE(E [Z_{T+h}|F_t]),$$

en el sentido que $MSE(\hat{\mathcal{Z}}_{T+h}) - MSE(E [Z_{T+h}|F_t]) \geq 0$, es decir, la matriz $MSE(\hat{\mathcal{Z}}_{T+h}) - MSE(E [Z_{T+h}|F_t])$ es definida no-negativa ver libro [16] Capítulo 7. Se puede verificar que se cumple la igualdad si $\hat{\mathcal{Z}}_{T+h} = E [Z_{T+h}|F_t]$. Note que si consideramos el proceso lineal 10.4 pero sólo para el subíndice mayor o igual a cero y asumiendo que $\{Z_t\} \sim i.i.d(\mathbf{0}, \Sigma_Z)$ tenemos que

$$E [X_{T+h}|F_t] = \mu + \mathbf{C}_h Z_T + \mathbf{C}_{h+1} Z_{T-1} + \cdots +$$

ya que $F_t = \{X_t, X_{t-1}, \dots\} = \{Z_t, Z_{t-1}, \dots\}$ así el error de predicción h -pasos adelante es

$$e_T(h) = X_{T+h} - E[X_{T+h}|F_t] = C_0 + C_1 Z_{T+h-1} + \dots + C_{h-1} Z_{T+1}$$

y la matriz de covarianza del error está dada por

$$Cov[e_T(h)] = \Sigma_Z + \sum_{i=1}^{h-1} C_i \Sigma_Z C_i'.$$

Ver libro [37] página 17 para detalles de los intervalos de predicción.

10.3 Regresión de Series de Tiempo

Antes de hablar de los modelos con respuesta multivariada, consideraremos modelos con respuesta univariada pero que pueden explicados por una o mas variables en un contexto de series de tiempo. Por ejemplo, se desea obtener el pronóstico de las ventas mensuales y con base en la inversión en publicidad x . Otro ejemplo consiste en predecir la demanda de electricidad y usando la temperatura x_1 y el día de la semana x_2 como predictores. Inicialmente consideraremos que las series de tiempo son estacionarias y que sólo hay una única variable explicativa estocástica, ya que para series no estacionarias nos podemos encontrar con relaciones espurias. Asimismo, tendremos modelos de regresión con errores tipo ARMA o ARIMA. Los modelos finales hacen referencia a los modelos de función de transferencia los cuales permiten incluir retardos de las variables predictoras y de la misma variable de respuesta.

10.3.1 Regresión Lineal simple

Para este modelo, vamos a considerar que las variables son del tipo determinísticas del tipo seno y coseno o dummy(para incluir estacionalidades o intervenciones).

Inicialmente vamos a considerar que tenemos dos series de tiempo estacionarias $\{y_t\}$ y $\{x_t\}$, tal que están relacionadas mediante la regresión:

$$y_t = \alpha + \beta x_t + e_t, \quad (10.6)$$

donde vamos a suponer que $\{e_t\} \sim RBG(0, \sigma^2)$. El supuesto de ser ruido blanco puede ser bastante restrictivo pero será considerado mas adelante. Bajo el supuesto hecho para los errores, tenemos que se pueden estimar los

parámetros del modelo usando mínimos cuadrados ordinarios(OLS). Veamos un ejemplo que aparece en el libro [19] que permite relacionar los cambios porcentuales trimestrales del gasto real del consumo personal(y) y el ingreso disponible real personal. El modelo de regresión lineal no se ajustó bien, por lo tanto consideraremos un modelo de regresión múltiple.

10.3.2 Regresión Lineal Múltiple

Por supuesto en varias ocasiones, no es suficiente una única variable para explicar la dinámica de otra variable. Por lo tanto, es necesario incluir otras variables al modelo, lo cual permite tener una mejor capacidad explicativa. Consideremos que se tienen las series $\{y_t\}$, $\{x_{1,t}\}, \dots, \{x_{k,t}\}$ y el modelo de regresión es:

$$y_t = \beta_0 + \beta_1 x_{1,t} + \dots + \beta_k x_{k,t} + e_t$$

donde $\{e_t\}$ tiene los mismos supuestos que en el caso de la regresión lineal simple. Bajo esas condiciones, también puede usarse OLS para estimar los parámetros del modelo. Vamos a considerar el mismo ejemplo de la regresión lineal simple pero ahora añadiremos mas predictores(Ingreso, producción, desempleo y ahorros).

La predicción de los modelos de regresión se hacen directamente como en la la regresión I.I.D., es decir

$$\hat{y}_t = \hat{\beta}_0 + \hat{\beta}_1 x_{1,t} + \dots + \hat{\beta}_k x_{k,t}.$$

Es importante verificar que los supuestos del modelo, ver sección 7.3 del libro digital [19]. Vamos a enfocarnos en el ACF y PACF de los residuales. Note que no hay una estructura de autocorrelación muy fuerte.

Nota 10.15. *Note que si en los residuales del modelo se encuentra estructura de autocorrelación temporal,y esa autocorrelación es de la familia ARMA(p, q), entonces es adecuado usar mínimos cuadrados generalizados, donde el estimador queda expresado:*

$$\hat{\beta} = (X' \Omega^{-1} X)^{-1} X' \Omega^{-1} y$$

donde $\Omega = \text{Cov}(\epsilon|X)$, y puede usarse nlme :: gls. Esta estrategia es un poco eficiente porque si la serie es larga, producir la inversa de la matriz de covarianza puede ser algo complejo.

10.3.3 Relaciones Espurias

Cuando las series que deseamos relacionar no son estacionarias, como sucede en muchas series de tiempo, el impacto pueden ser muy grande cuando se desea llevara a cabo una regresión lineal. Puede suceder que aunque dos series de tiempo sean independientes pero no estacionarias, al llevar a cabo una regresión entre ellas, la regresión nos muestre que si hay dependencia entre ellas. Eso lo conocemos como relaciones espurias. Veamos una simulación en R.

10.3.4 Otros Regresores útiles

Hay mucho regresores útiles que pueden ser usados cuando se hace regresión para series de tiempo. Ejemplos de estos son:

- **El Tiempo como Predictor.** El tiempo puede ser un predictor, cuando deseamos modelar una tendencia. Por ejemplo una tendencia lineal puede modelarse estableciendo $x_{1,t} = t$, y así

$$y_t = \beta_0 + \beta_1 t + \epsilon_t.$$

Si se usa la función `fable::TSLM` podemos incluir este tipo de tendencias a través de la función `trend()` que función dentro de este entorno.

- **Variables Dummy.** Cuando un predictor es una variable categórica, como por ejemplo, para el caso de si un día es festivo o no. Es posible indicar esto mediante una variable dummy o una variable indicadora. Recuerden que en el análisis de intervención o de outliers se involucran este tipo de variables. Si hay mas de dos categorías, entonces la variable debe ser codificada por varias variables dummy(una menos que el número total de categorías). La función `fable::TSLM` manipula estos caso si se especifica una variable factor como el predictor.
- **Variables Dummy Estacionales.** Si se ha detectado una estacionalidad de periodo s , esta puede ser modelada mediante variable dummy como predictores. supongamos que se tiene un periodo $s = 7$ para una serie diaria. Entonces las siguientes variables dummy deben ser creadas como se especifica en la tabla 10.3.4:

	$D_{1,t}$	$D_{2,t}$	$D_{3,t}$	$D_{4,t}$	$D_{5,t}$	$D_{6,t}$
Lunes	1	0	0	0	0	0
Martes	0	1	0	0	0	0
Miércoles	0	0	1	0	0	0
Jueves	0	0	0	1	0	0
Viernes	0	0	0	0	1	0
Sábado	0	0	0	0	0	1
Domingo	0	0	0	0	0	0
Lunes	1	0	0	0	0	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮

Tabla 10.1: Tabla efectos variables dummy por día

Vale la pena decir que si al modelo se le incluye el intercepto, requiere una variable dummy menos, ya que el efecto de la variable restante es capturado por dicho intercepto, el cual es obtenido estableciendo todas las variables dummy en cero. La función `fable::TSLM` junto `season()` manipulan este tipo de efectos.

- **Variables de intervención.** Las variables de intervención son útiles para cuando la ocurrencia de un evento hace que la dinámica de la variable de interés cambie con respecto a la dinámica presentada en el histórico de la serie. Por ejemplo la aparición del coronavirus hizo que por ejemplo el desempleo en Colombia aumentara pero con el paso de los meses tiende a volver a los niveles pre-pandemia, esa es una intervención del tipo transitoria ver figura 10.8.

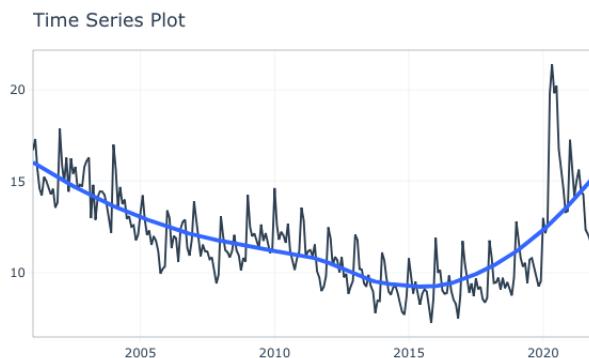


Figura 10.8: Intervención Desempleo debido al Covid.

En este caso la variable que debería ser incluida tiene la siguiente forma:

Fecha	I_t
⋮	⋮
2020-01	0
2020-02	0
2020-03	0
2020-04	δ^0
2020-05	δ^1
2020-06	δ^2
2020-07	δ^4
2020-08	δ^5
⋮	⋮

Tabla 10.2: Intervención transitoria Covid sobre el Desempleo

Por supuesto, dependiendo de como sea el efecto de la intervención sobre la variable de interés, así mismo será la variable de intervención que será incluida. Note que estas variable de intervención son usadas para modelar outliers como se vio en el capítulo 8.

- **Días Laborales.** El número de días laborales en el mes puede tener un efecto sobre las ventas por ejemplo. Entonces incluir una variable que indique el número de días laborales en el mes puede mejorar el pronostico.
- **Rezagos Distribuidos.** En ocasiones, retardos de las variables predictoras pueden afectar el valor presente de la variable interés, por eso esas variable pueden incluidas en el modelo. Ese tipo de modelos serán parte de una exposición.
- **Semana Santa.** En algunos países, la celebración de la semana santa puede tener impacto sobre las variables de interés, por lo tanto es necesario tener en cuenta esto, y mas aún porque en cada año el mes en que se celebra la semana santa puede diferir. Es por esto que una variable dummy que indique el mes en que celebra la semana santa en cada año, puede ser importante. Es importante señalar que a veces la semana santa tomas días de marzo y abril, lo cual implica que hay que dividirse proporcionalmente entre los meses.

- **Serie de Fourier.** Una alternativa a las variables dummy para modelar estacionalidades, especialmente para periodos estacionales largos, consiste en usar términos o predictores de Fourier como se explicó en 2.8 capítulo 2. Por ejemplo si tenemos que el periodo estacional es s , entonces los primeros términos de Fourier están dados por:

$$x_{1,t} = \sin\left(\frac{1 \cdot 2\pi t}{s}\right), \quad x_{2,t} = \cos\left(\frac{1 \cdot 2\pi t}{s}\right), \quad x_{3,t} = \sin\left(\frac{2 \cdot 2\pi t}{s}\right),$$

$$x_{4,t} = \sin\left(\frac{2 \cdot 2\pi t}{s}\right), \quad x_{5,t} = \cos\left(\frac{3 \cdot 2\pi t}{s}\right), \quad x_{6,t} = \sin\left(\frac{3 \cdot 2\pi t}{s}\right),$$

donde la expansión de N términos está dada por:

$$\sum_{i=1}^N \left(a_n \sin\left(\frac{i \cdot 2\pi t}{s}\right) + b_n \cos\left(\frac{i \cdot 2\pi t}{s}\right) \right)$$

Si tenemos una estacionalidad de periodo $s = 12$, entonces idealmente deberíamos usar una valor N mas pequeño que s , el cual es usualmente un valor de máximo $N = s/2$. Se le llama regresión armónica si sólo son utilizados los dos primeros términos $x_{1,t}$ y $x_{2,t}$.

Hoy en día, en el lenguaje del aprendizaje automático, la elección y construcción de las variables predictoras se conoce como *ingeniería de características*, y del cual hablaremos un poco mas adelante. Bibliografía relacionada con esto y que menciona el contexto de series de tiempo se puede encontrar en [7] y [26].

10.3.5 Selección de predictores

Cuando se tienen varios predictores, se requiere de una estrategia para seleccionar los mejores predictores que serán usado en la regresión. Una forma de hacer estos es usando una medida de la precisión predictiva. Existen varias medidas, entre estas están: AIC , $AICc$, BIC , $CV(MSE)$, R^2 .

10.3.6 Pronósticos con modelos de Regresión

Por supuesto, un modelo de regresión puede ser usado para obtener pronósticos, es decir obtener el valor que va tomar la variable aleatoria de interés en los tiempos subsecuentes o futuros. Sin embargo hay dos tiempos de pronósticos.

Pronósticos Ex-ante Consideremos que hemos observado valores en los tiempos $t = 1, \dots, T$ para la variable de respuesta Y_t , y las covariables o variables predictoras $x_{i,t}$ para $i = 1, \dots, k$. Se desea predecir los valores futuros de $Y_{T+1}, Y_{T+2}, \dots, Y_{T+h}$ con base únicamente en la información $x_{i,1:T}$ para todo $i = 1, \dots, k$. Es decir, hay que tener los pronósticos de los predictores.

Pronóstico Ex-post Los pronósticos son realizadas utilizando información posterior sobre los predictores, es decir, hay que usar las verdaderas observaciones de los predictores en el modelo. Esas no son genuinos pronósticos, pero son útiles para evaluar los modelos de pronóstico.

Intervalos de Predicción Si tenemos una regresión lineal simple tal que el modelo ajustado es de la forma

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x,$$

y asumiendo que los errores tiene distribución normal, un intervalos de predicción para y viene dado por:

$$\hat{y} \pm 1.96 \hat{\sigma}_\epsilon \sqrt{1 + \frac{1}{T} + \frac{(x - \bar{x})^2}{(T-1)s_x^2}}$$

con T el tamaño de la muestra, \bar{x} la media de los valores observados de la variable x , s_x es su respectiva desviación estándar, mientras que

$$\hat{\sigma}_\epsilon^2 = \frac{1}{T-k-1} \sum_{t=1}^T e_t^2$$

donde $e_t = y_t - \hat{y}_t$.

Pronósticos Basados en Escenarios

Supongamos que la persona que produce los pronósticos asume escenarios posibles para las variables predictoras que son de interés. Por ejemplo, se está interesado en comparar el cambio predicho en el consumo cuando hay un crecimiento constante de 1% y 0.5% respectivamente para el ingreso y los ahorros, sin cambios en la tasa de desempleo, versus una caída respectiva de 1% y 0.5%, para cada uno de los trimestres luego del final del periodo de la muestra. Note que los intervalos de predicción calculados no tienen en cuenta la incertidumbre asociada a los valores futuros de las variables predictoras.

Construyendo un modelo de regresión predictivo Por supuesto, uno desea en la práctica producir pronósticos h - pasos adelante. Es decir, para un pronóstico ex-ante se requieren los valores futuros de cada predictor. Pero obtener los pronósticos de pronósticos puede ser complicado, por lo tanto

se requieren alternativas para hacer los pronósticos de la variable de interés. La primera alternativa es usar pronósticos basados en escenarios como se explicó anteriormente. Otra alternativa práctica consiste en que el valor futuro h -pasos adelante de la variable de interés, depende de los valores rezagados de sus predictores, es decir, consideramos que le modelo es de la siguiente forma:

$$y_{t+h} = \beta_0 + \beta_1 x_{1,t} + \cdots + \beta_k x_{k,t} + e_{t+h}$$

para $h = 1, 2, \dots$

Leer sección 7.8 libro [19] acerca de correlación, y el pronóstico.

10.4 Modelos de Regresión Dinámica

Los modelos regresión que se consideraron anteriormente

$$y_t = \beta_0 + \beta_1 x_{1,t} + \cdots + \beta_k x_{k,t} + \epsilon_t$$

consideraban que $\{\epsilon_t\} \sim RB(0, \sigma^2)$. Sin embargo este modelo puede ser bastante restrictivo y se planteará el escenario donde los errores tengan estructura de autocorrelación. Vamos a considerar que los errores del modelo de regresión son $\{\eta_t\}$ tal que

$$y_t = \beta_0 + \beta_1 x_{1,t} + \cdots + \beta_k x_{k,t} + \eta_t$$

y que asumimos que el modelo de los errores es un *arima*(p, d, q)

$$\phi(B)(1 - B)^d \eta_t = \theta(B)\epsilon_t$$

donde $\{\epsilon_t\} \sim RB(0, \sigma^2)$, es decir, el modelo tiene dos errores, los errores de la regresión η_t , y el error del modelo *arima* que es ϵ_t que es un ruido blanco. Durante el proceso de estimación se requiere minimizar la suma de cuadrados de los errores ϵ_t y no de η_t , por qué? Con estos, también, se puede usar máxima verosimilitud como proceso de estimación alternativa.

Nota 10.16. *Cuando se lleva a cabo el proceso de estimación, se deben considerar al menos dos escenarios:*

- 1) *Se desea estimar la regresión con errores $\{\eta_t\} \sim ARMA$. Por lo tanto todas las variables deben ser estacionarias, es decir tanto la variable de respuesta y como las predictoras. La excepción a esto es que las variables no estacionarias deberían estar cointegradas, es decir, hay una combinación lineal de las variables que es estacionaria ($y_t, x_{1,t}, \dots, x_{k,t}$).*

- 2) Si algunas o todas las variables son no estacionarias y no hay relaciones de cointegración, la sugerencia es tome diferencia ordinaria a todas las variables y lleve a cabo la regresión con las series diferenciadas, este modelo se llama **modelo en diferencias**, distinto al **modelo en niveles** que obtenido en un modelo sin diferenciar las series. Eso puede romper las estructuras de relación entre las variables.
- 3) Se puede verificar que un modelo de regresión con errores arima es equivalente a un modelo de regresión en diferencias con errores arma, ver ejemplo sección 10.1 del libro [19].

10.4.1 Pronóstico

Para generar los pronósticos del modelo de regresión con errores ARIMA se requiere obtener el pronóstico de la parte de la regresión y el pronóstico de la parte del modelo **ARIMA**, para luego combinar los resultados. Por supuesto deben conocerse los valores futuros de las variables predictoras o hacer un modelamiento de ellas de forma separada y obtener los pronósticos.

10.4.2 Predictores retardados

Es frecuente encontrar que un predictor no tiene efecto instantáneo sobre la variable de respuesta. Por ejemplo pensemos en que la disminución en la ingesta de calorías en mes determinado no tiene efecto inmediato sobre el peso, sino que un mes o dos mese después se ve la disminución en el peso. Otro ejemplo consiste en que el incremento en publicidad en un mes tiene efecto sobre las ventas al siguiente mes o inclusive varios mese después, es decir el efecto no es inmediato sobre la variable de respuesta cuando hay un cambio en la variable predictora. En este caso, es necesario permitir efectos retardados de la variable predictora . Vamos a suponer por el momento que hay sola una variable predictora, entonces el modelo con efectos retardados queda de la siguiente manera:

$$y_t = \beta_0 + \gamma_0 x_t + \gamma_1 x_{t-1} + \cdots + \gamma_k x_{t-k} + \eta_t$$

donde $\{\eta_t\} \sim ARIMA$, uno puede explorar el valor de k usando la función de autocorrelación cruzada. También puede elegirse el valor de k junto con el de p, q talque se minimice algún criterio de información o vía validación cruzada tal que minimice un criterio de pronóstico.

10.5 Árboles de Regresión y Bosques Aleatorios

Esta sección está basa en el libro de [30].

Los árboles son modelos simples basados en particiones recurrentes para explorar la estructura de los datos. Tales modelos son usados en un contexto de clasificación y de regresión. La idea en un contexto de predicción consiste esencialmente en la partición del espacio predictor en sub-regiones que no se traslanan y de la media muestral de la variable dependiente en cada región como su predicción. Se puede pensar un árbol de regresión con múltiples predictores como una función de salto multivariada. La pregunta clave en este contexto consiste en encontrar formas eficientes de particionar el espacio predictor talque se minimice la suma de cuadrados del error de pronóstico en la muestra de entrenamiento.

Se va a iniciar con los árboles binarios y las formas en que crece y se poda. Usualmente el tamaño de un árbol consiste en el número de particiones, o el número de sub-regiones resultantes del espacio predictor. Hay dos casos extremos, el primero consiste en que no hay partición del espacio predictor, mientras que en contraste tenemos el árbol mas grande que se puede conformar, el cual consiste en que cada punto de los datos resulta ser una hoja. En el caso del árbol mas pequeño la predicción resulta ser la media de todas las observaciones de la variable de respuesta, mientras que en el caso del árbol mas grande se usa como predicción un solo dato. Por lo tanto, ni el árbol mas pequeño, ni el árbol mas grande resulta útil en la práctica. Así que lo métodos basado en árboles son útiles si se logra construir un árbol donde la predicción es hecha usando la media condicional de la variable de respuesta, tal que las variables explicativas toman en una región definida del espacio predictor. Para esto será útil definir el crecimiento y la poda del árbol.

10.5.1 Crecimiento

Considere los datos $\{(\underline{x}_t, y_t) | t = 1, \dots, n\}$, donde $\underline{x}_t = (x_{1t}, \dots, x_{kt})'$ es una realización del predictor \underline{X}_t , y y_t es una realización de la variable dependiente o de respuesta Y_t con n denotando el tamaño de la muestra. Un árbol de regresión es típicamente un árbol binario construido por el crecimiento y poda de ramas que permitan describir las relaciones entre \underline{X}_t y Y_t de una mejor forma. Las hojas son llamados los nodos terminales y cada rama da una partición sobre el espacio de predictores. Vamos a considerar la

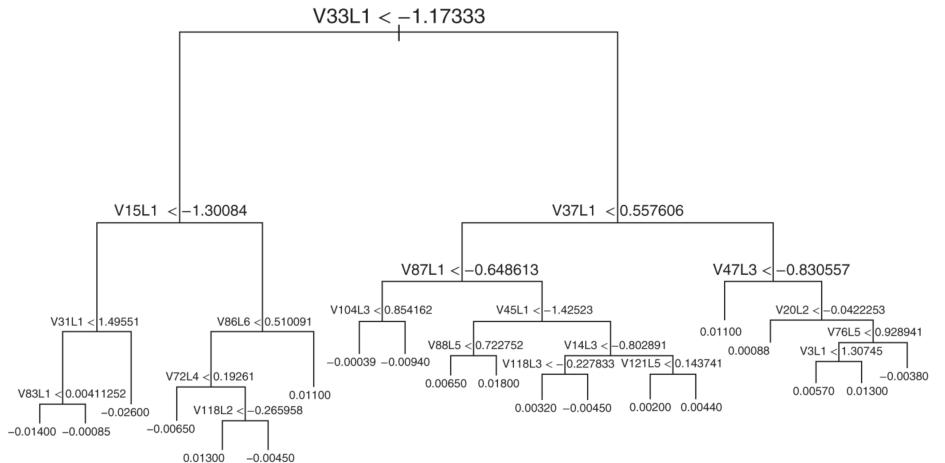


Figura 10.9: Árbol para predecir la tasa de crecimiento de producción industrial US

salida 10.9 del ejemplo 8.1 del libro [30] para explicar un poco los conceptos principales. En el tallo, el espacio predictor es \mathbb{R}^k . Es decir, no hay partición del espacio. Para el i -ésimo predictor x_{it} , sean los estadísticos de orden $x_{i(1)} \leq x_{i(2)} \leq \dots \leq x_{i(n)}$. Para cualquier número real $\eta \in [x_{i(1)}, x_{i(n)}]$, particiona \mathbb{R}^k en las siguientes dos sub-regiones:

$$R_{i,1} = \{\underline{x}_t | x_{it} \leq \eta\}, \quad R_{i,2} = \{\underline{x}_t | x_{it} > \eta\}.$$

De aquí se puede computar la suma de cuadrados del residual y_t de la partición como

$$SS(i, \eta) = \sum_{\underline{x}_t \in R_{i,1}} (y_t - \bar{y}_1)^2 + \sum_{\underline{x}_t \in R_{i,2}} (y_t - \bar{y}_2)^2,$$

donde \bar{y}_1 y \bar{y}_2 son las medias muestrales de y_t en $R_{i,1}$ y $R_{i,2}$ respectivamente. En la práctica, η asume el estadístico de orden de x_{it} y la mejor partición se selecciona de tal manera que

$$\hat{\eta}_i = \arg \min_{\eta \in \{x_{i(1)}, \dots, x_{i(n)}\}} SS(i, \eta),$$

donde el subíndice i se adiciona para hacer énfasis en el i -ésimo predictor. La suma de cuadrados del error de pronóstico resultante $SS(i, \hat{\eta}_i)$. Para determinar la primera rama se escoge el predictor i_1 con el respectivo

umbral η_{i_1} , donde i_1 es dado por:

$$(i_1, \hat{\eta}_{i_1}) \arg \min_{1 \leq i \leq k} SS(i, , \hat{\eta}_i),$$

y las sub-regiones asociadas son $R_{i_1,1}$ y $R_{i_1,2}$, con lo cual la primera partición termina y con dos hojas, es decir, $R_{i_1,1}$ y $R_{i_1,2}$.

El proceso de crecimiento es repetido para $R_{i_1,1}$ y $R_{i_1,2}$ de forma separada. La única modificación para $R_{i_1,1}$ es que $x_{i_1,t}$ solo asume valores menores o iguales a $\hat{\eta}_{i_1}$. Para $R_{i_1,2}$, $x_{i_1,t}$, se asume valores mas grandes que $\hat{\eta}_{i_1}$. Note que el proceso de crecimiento puede continuar hasta que cada sub-región contenga una sola observación o un número pequeño de datos pre-especificado. Si se deja crecer hasta este punto, nos encontraremos en un problema de sobreajuste. Para evitar el sobre ajuste se requiere de un método llamado la poda de árbol.

10.5.2 Poda del Árbol

El número de divisiones de un árbol de regresión binaria es llamada la *profundidad* del árbol. Para profundidades mayores a cero, el número de hojas(o nodos terminales) es igual a la profundidad del árbol más 1, y se le conoce como *tamaño* del árbol. Sea T_m un árbol de profundidad m . Este árbol divide el espacio de predictores \mathbb{R}^k dentro de $m + 1$ sub-regiones denotadas $\{R_j | j = 1, \dots, m+1\}$, donde por simplicidad no se muestran los predictores incluidos, ver libro [38] capítulo 4 para detalles. La suma de cuadrados de los errores de pronósticos para el árbol T_m se puede escribir como:

$$SS(T_m) = \sum_{j=1}^{m+1} \sum_{x_t \in R_j} (y_t - \bar{y}_j)^2,$$

donde \bar{y}_j denota la media muestral de y_t en R_j , mas específicamente es la media condicional observada de la respuesta cuando las variables explicativas pertenecen a R_j . La poda del árbol emplea por lo general el criterio del costo de la complejidad dado por

$$C(T_m, \lambda) = SS(T_m) + \lambda|T_m|,$$

con $|T_m|$ siendo el tamaño del árbol y $\lambda > 0$ es el parámetro de penalidad, el cual gobierna la compensación entre la complejidad del árbol y la bondad del ajuste del mismo. En la práctica λ puede encontrarse vía validación cruzada o validando los pronósticos fuera de la muestra. Así, para un valor

de λ fijo, se selecciona el árbol tal que

$$T_{\hat{m}} = \arg \min_m C(T_m, \lambda).$$

Se pueden usar los paquete *tree* o *rpart* para llevara a cabo la regresión basados en árboles.

Ejemplo 10.17. Vamos a considerar que la variable de interés es la producción industrial, la cual depende de otras 122 variables. Esos datos están descrito en McCracken y Ng (2016) y disponible en <https://research.stlouisfed.org/econ/mccracken/fred-databases/>. El conjunto de datos se actualiza mensualmente y contiene 128 variables macroeconómicas. Los detalles de las variables se dan en McCracken y Ng (2016). Debido a los valores faltantes, eliminamos seis variables (numeradas 58, 60, 95, 104, 123 y 128) y usamos los datos desde enero de 1960 hasta abril de 2019 en nuestro análisis. Además, usamos la diferenciación y la transformación dadas en McCracken y Ng (2016) para crear series estacionarias.

10.5.3 Bosques Aleatorios

Para poder hablar de bosques aleatorios se requiere de la idea de ensamblaje. La idea que hay detrás los métodos de ensamblaje, consiste en combinar diferentes procedimientos de pronóstico o algoritmos, para obtener una mejor predicción que la obtenida individualmente por cada algoritmo de forma separada, como sucede en *model averaging*. Los bosques aleatorios son una extensión del bagging(Agregación bootstrap) para evitar el sobreajuste en los métodos de predicción basados en árboles. La idea detrás de los árboles consiste en emplear muchos árboles simples para producir predicciones, luego combinar aquellas predicciones para producir un pronóstico en consenso. Este modelo de árboles se considera no paramétrico porque no asume un modelo específico subyacente para los datos.

10.5.4 Bagging

Vamos a introducir la idea del bagging inicialmente. Consideremos que el conjunto de datos $\{(\underline{x}_t, y_t) | t = 1, \dots, n\}$, donde \underline{x}_t es una realización del predictor $\underline{X}_t = (X_{1t}, \dots, X_{kt})'$ y y_t es una realización de la variable dependiente o de respuesta Y_t . El objetivo es predecir y_{t+h} , con h siendo el horizonte de pronóstico. Entonces el bagging consiste de:

1. Extraiga una muestra aleatoria con reemplazo del conjunto de datos. Denote la muestra bootstrap como $\{(x_t^*, y_t^*)|t = 1, \dots, n\}$. **Nota:** Observe que no se debería usar el bootstrap para muestras i.i.d ya que estamos en un contexto de muestras autocorrelacionadas.
2. Construya un árbol de regresión usando la muestra bootstrap para obtener la predicción y_{t+h} .

Ahora, el bosque aleatorio extiende el procedimiento bagging introduciendo la idea de explorar completamente la relación entre y_{t+h} y x_t . La idea es que, para cada división del árbol, se selecciona una muestra aleatoria de los predictores como candidatos para la partición. Sea g un entero positivo tal que $1 \leq g \leq k$, usualmente $g = [k/3]$. Para construir un árbol de regresión de una muestra bootstrap para el bosque aleatorio, uno toma una muestra de g predictores de $\{1, \dots, k\}$ que sirvan para la partición en cada división. Este paso nuevo sirve para múltiples propósitos. Primero, esto reduce el cómputo que debe realizarse, en especial cuando k es grande. Segundo, esto mitiga la multicolinealidad, porque los predictores altamente correlacionados podrían no ser seleccionados juntos. Finalmente, se evita el chance de usar árboles similares en el ensamblaje. El promedio simple de los pronósticos de cada árbol es usado para producir un pronóstico consensuado. Un promedio ponderado puede ser útil dependiendo la función de pérdida seleccionada.

La siguiente discusión es sacada del libro [30].

Nota 10.18. *Algunas discusiones sobre RF están en orden. Primero, el paquete R **randomForest** se puede usar para realizar predicciones de RF. En segundo lugar, se necesitan algunas modificaciones para aplicar RF de manera efectiva a los datos dependientes. Por ejemplo, en el análisis de series temporales, el valor de retraso 1 de la variable dependiente suele ser más importante que el valor de retraso 2. De manera similar, los retrasos estacionales de la variable dependiente suelen ser importantes en el modelo de series temporales estacionales. En consecuencia, se podría usar una muestra aleatoria ponderada para seleccionar los predictores como candidatos para dividir en cada división en lugar de usar una muestra aleatoria pura. En tercer lugar, se debe tener cuidado al extraer muestras de arranque para datos dependientes y se debe usar alguna forma de arranque en bloque; véase, por ejemplo, Alonso et al. (2004) y Lahiri (2013).*

En cuanto al tema de boostrap pido leer el artículo RANDOM FORESTS FOR TIME-DEPENDENT PROCESSES de Benjamin Goehry y su posterior estudio de simulación y con datos reales del artículo <https://hal>.

archives-ouvertes.fr/hal-03129751/document donde usa el paquete *rangerts* cuya implementación está en <https://github.com/hyanworkspace/rangerts>.

10.6 Aspectos a tener en cuenta en un modelo de regresión

Note que en el modelamiento, en un contexto de aprendizaje supervisado, se deben explorar qué características presenta la variable de repuesta o de interés, por ejemplo si la serie asociada a esta variable presenta estacionalidad o varias estacionalidades, si se presenta, efectos de calendario (semana santa, acción de gracias, super-bowl, días laborales y no laborales, etc), si requiere transformación Box-Cox. Sin embargo, se debe tener en cuenta también los siguientes aspectos:

1. La inclusión o no de retardos de las variables predictoras.
2. Exploración de relaciones no lineales entre las variables.
3. Si las variables deben entrar en niveles o en diferencias, o des-estacionalizadas.
4. Cómo seleccionar las variables predictoras.

Estos y otros aspectos deben tenerse en cuenta en el modelamiento. En especial, si tienen muchas variables predictoras y se desea hacer una regresión lineal, vale la pena tener en modelos de regresión con regularización o hacer un paso previo el cual es hacer componentes principales. Sin embargo en la literatura [15], [25], [7] se presenta un discusión acerca que de métodos para seleccionar variables en el modelamiento predictivo, lo cuales pueden ser aplicados en el contexto de series de tiempo como por ejemplo métodos *Wrapper* y *Embedded*.

11

Procesos VARMA

En este capítulo introduciremos la generalización al caso multivariado de los procesos autorregresivos, o mas conocidos como los vectores autorregresivo(VAR), los cuales permiten modelar múltiples series de tiempo en el sentido que varias variables de interés son observadas en un instante del tiempo t . Los modelos utilizados inicialmente se emplean para modelar series de tiempo de baja dimensión. Vamos a seguir mayormente el libro [27], sin embargo libros complementarios son [30], [41] y [37].

11.1 Supuestos Básicos y propiedades de los Procesos VAR

11.1.1 Procesos Estables VAR(p)

Veamos ahora la definición del proceso VAR(p) o también llamado modelo vectorial autorregresivo de orden p

Definición 11.1. *Un proceso estocástico K -dimensional $\{\underline{Y}_t\}$ se dice que sigue un proceso VAR(P) si es solución de la ecuación en diferencias estocástica*

$$\underline{Y}_t = \underline{\nu} + A_1 \underline{Y}_{t-1} + \cdots + A_p \underline{Y}_{t-p} + \underline{u}_t, \quad t = 0, \pm 1, \pm 2, \dots,$$

donde A_i son matrices $K \times K$ de coeficientes fijos, $\underline{\nu} = (\nu_1, \dots, \nu_K)'$ es el vector de interceptos. El proceso $\{\underline{u}_t = (u_{1t}, \dots, u_{Kt})'\}$ es un proceso ruido blanco K -dimensional, es decir, $E[\underline{u}_t] = 0$ y $E[\underline{u}_t \underline{u}_t'] = \Sigma_u$ y $E[\underline{u}_t \underline{u}_s'] = 0$ para $s \neq t$. La matriz de covarianza Σ_u se asume no singular(es decir si

posee inversa), y por lo general es una matriz diagonal, para que no hayan interrelaciones entre los ruidos.

Para ganar un poco mas de conocimiento acerca de los procesos VAR, vamos a estudiar un poco el proceso VAR(1). Sea $\{\underline{Y}_t\}$ un proceso VAR(1) tal que

$$\underline{Y}_t = \nu + A_1 \underline{Y}_{t-1} + \underline{u}_t,$$

de tal forma que vamos a iterar t veces el proceso hasta el tiempo t .

$$\begin{aligned} y_1 &= \nu + A_1 y_0 + u_1, \\ y_2 &= \nu + A_1 y_1 + u_2 = \nu + A_1(\nu + A_1 y_0 + u_1) + u_2 \\ &= (I_K + A_1)\nu + A_1^2 y_0 + A_1 u_1 + u_2, \\ &\vdots \\ y_t &= (I_K + A_1 + \cdots + A_1^{t-1})\nu + A_1^t y_0 + \sum_{i=0}^{t-1} A_1^i u_{t-i} \\ &\vdots \end{aligned}$$

Así, note que el vector \underline{y}_t está determinado únicamente por $\underline{y}_0, \underline{u}_1, \dots, \underline{u}_t$, por lo que la distribución conjunta de $\underline{y}_1, \dots, \underline{y}_t$ está determinada por la distribución conjunta de $\underline{y}_0, \underline{u}_1, \dots, \underline{u}_t$. Qué sucede si no se inicia el proceso en $t = 0$, sino que empieza en el pasado infinito? Esto nos permite que si para un tiempo t fijo, escribimos la recurrencia hasta el $j + 1$ paso, tenemos que

$$\begin{aligned} \underline{Y}_t &= \nu + A_1 \underline{Y}_{t-1} + u_t \\ &= (I_K + A_1 + \cdots + A_1^j)\nu + A_1^{j+1} \underline{Y}_{t-j-1} + \sum_{i=0}^j A_1^i u_{t-i}. \end{aligned}$$

Usando el resultado de la página 657 del libro [27], que nos dice que si todos los valores propios de A_1 tienen módulo mas pequeño que 1, entonces la sucesión $\{A_1^j, i = 0, 1, \dots\}$ es absolutamente sumable, y la razón básicamente es la descomposición canónica de Jordan de la matriz $A_1 = P\Lambda P^{-1}$, y así $A_1^j = (P\Lambda P^{-1})^j = P\Lambda^j P^{-1}$, con

$$\Lambda = P^{-1} A_1 P = \begin{bmatrix} \Lambda_1 & & 0 \\ & \ddots & \\ 0 & & \Lambda_K \end{bmatrix}$$

donde cada matriz Λ_i para $i = 1, \dots, K$ es triangular superior

$$\Lambda_i = \begin{bmatrix} \lambda_i & 1 & 0 & \cdots & 0 \\ 0 & \lambda_i & 1 & & 0 \\ \vdots & & \ddots & \ddots & \vdots \\ 0 & 0 & & \ddots & 1 \\ 0 & 0 & \cdots & \cdots & \lambda_i \end{bmatrix}$$

Mas específicamente se tiene lo siguiente:

Rules: Suppose A is a real $(m \times m)$ matrix with eigenvalues $\lambda_1, \dots, \lambda_n$ which have all modulus less than 1, that is, $|\lambda_i| < 1$ for $i = 1, \dots, n$. Furthermore, let Λ and P be the matrices given in (A.9.1).

- (1) $A^j = PA^jP^{-1} \xrightarrow{j \rightarrow \infty} 0$.
- (2) $\sum_{j=0}^{\infty} A^j = (I_m - A)^{-1}$ exists.
- (3) The sequence A^j , $j = 0, 1, 2, \dots$, is absolutely summable, that is,

$$\sum_{j=0}^{\infty} |\alpha_{kl,j}|$$

is finite for all $k, l = 1, \dots, m$, where $\alpha_{kl,j}$ is a typical element of A^j . (See Section C.3 regarding the concept of absolute summability.)

Con lo cual se puede verificar que

$$\sum_{i=0}^{\infty} A_1^i \underline{u}_{t-i}$$

es convergente en media cuadrática. También, se puede verificar que

$$(I_K + A_1 + \cdots + A_1^j) \underline{\nu} \rightarrow_{j \rightarrow \infty} (I_k - A_1)^{-1} \underline{\nu}.$$

Finalmente, note que como $A_1^{j+1} \xrightarrow{j \rightarrow \infty} 0$ y asumiendo que $E[\underline{Y}'_{t-j} \underline{Y}_{t-j}] < \infty$ para cada j , por lo tanto

$$\begin{aligned} & \lim_{j \rightarrow \infty} E[(\underline{Y}_t - (I_K + A_1 + \cdots + A_1^j) \underline{\nu} - \sum_{i=0}^j A_1^i \underline{u}_{t-i})' (\underline{Y}_t - (I_K + A_1 + \cdots + A_1^j) \underline{\nu} - \sum_{i=0}^j A_1^i \underline{u}_{t-i})] \\ &= \lim_{j \rightarrow \infty} E[(A_1^{j+1} \underline{Y}_{t-j-1})' (A_1^{j+1} \underline{Y}_{t-j-1})] \rightarrow 0, \end{aligned}$$

entonces \underline{Y}_t está bien definido para cada t en media cuadrática, entonces

$$\underline{Y}_t = \mu + \sum_{i=0}^{\infty} A_1^i \underline{u}_{t-i},$$

con $\mu = (I_k - A_1)^{-1} \underline{\nu}$, es decir escritura como un proceso $MA(\infty)$. Note que las distribuciones del proceso $\{\underline{Y}_t\}$ depende únicamente de las

distribuciones del proceso $\{\underline{y}_t\}$. Basado en 10.11 y 10.12 de un proceso lineal, el vector de medias y matriz de covarianzas del proceso VAR(1) queda

$$E[\underline{Y}_t] = \mu \quad (11.1)$$

y

$$\Gamma_{\underline{Y}}(h) = Cov(\underline{Y}_t, \underline{Y}_{t-h}) = \sum_{i=0}^{\infty} A_1^{h+i} \Sigma_u A_1^{i'} \quad (11.2)$$

Definición 11.2. *El proceso VAR(1) se llama estable si todos los valores propios de A_1 tienen módulo más pequeño que 1. Esta condición es análoga a la condición de causabilidad de un proceso ARMA univariado.*

Nota 11.3. *Note que siguiendo la regla 7 de la página 653 del libro [27], es decir*

- (1) If A is symmetric, then all its eigenvalues are real numbers.
- (2) The eigenvalues of a diagonal matrix are its diagonal elements.
- (3) The eigenvalues of a triangular matrix are its diagonal elements.
- (4) An $(m \times m)$ matrix has at most m eigenvalues.
- (5) Let $\lambda_1, \dots, \lambda_m$ be the eigenvalues of the $(m \times m)$ matrix A , then $|A| = \lambda_1 \cdots \lambda_m$, that is, the determinant is the product of the eigenvalues.
- (6) Let λ_i and λ_j be *distinct* eigenvalues of A with associated eigenvectors v_i and v_j . Then v_i and v_j are linearly independent.
- (7) All eigenvalues of the $(m \times m)$ matrix A have modulus less than 1 if and only if $\det(I_m - Az) \neq 0$ for $|z| \leq 1$, that is, the polynomial $\det(I_m - Az)$ has no roots in and on the complex unit circle.

la condición de estabilidad para el proceso VAR(1) $\{\underline{Y}_t\}$, es que

$$\det(I_K - A_1 z) \neq 0 \text{ para } |z| \leq 1,$$

es decir, el polinomio característico no tiene raíces sobre el círculo unitario complejo, es decir sus raíces están por fuera del círculo unitario complejo. Vale la pena decir que aunque no se satisfaga la condición de estabilidad, el proceso puede estar bien definido.

Nota 11.4. *Note que decir que las raíces de*

$$\det(I_K - A_1 z) \neq 0 \text{ para } |z| \leq 1,$$

es equivalente a decir que los valores propios de A_1 están por dentro del círculo unitario, ya que son los inversos de las raíces de la ecuación anterior.

Nota 11.5. Note que el proceso $VAR(p)$

$$\underline{Y}_t = \underline{\nu} + A_1 \underline{Y}_{t-1} + \cdots + A_p \underline{Y}_{t-p} + u_t, \quad t = 0, \pm 1, \pm 2, \dots,$$

puede escribirse como un proceso $VAR(1)$ de mayor dimensionalidad, es decir, Kp -dimensional, en efecto, definiendo

$$Y_t := \begin{bmatrix} y_t \\ y_{t-1} \\ \vdots \\ y_{t-p+1} \end{bmatrix}_{(Kp \times 1)}, \quad \boldsymbol{\nu} := \begin{bmatrix} \nu \\ 0 \\ \vdots \\ 0 \end{bmatrix}_{(Kp \times 1)},$$

$$\mathbf{A} := \begin{bmatrix} A_1 & A_2 & \dots & A_{p-1} & A_p \\ I_K & 0 & \dots & 0 & 0 \\ 0 & I_K & & 0 & 0 \\ \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & I_K & 0 \end{bmatrix}_{(Kp \times Kp)}, \quad U_t := \begin{bmatrix} u_t \\ 0 \\ \vdots \\ 0 \end{bmatrix}_{(Kp \times 1)}.$$

entonces, $\underline{Y}_t = \underline{\nu} + \mathbf{A} \underline{Y}_{t-1} + \underline{U}_t$.

Por lo tanto, el proceso $VAR(P)$ es estable si

$$\det(I_{Kp} - \mathbf{A}z) \neq 0, \quad \text{para } |z| \leq 1.$$

Con lo cual la media y la función de autocovarianza matricial queda establecido como:

Its mean vector is

$$\boldsymbol{\mu} := E(Y_t) = (I_{Kp} - \mathbf{A})^{-1} \boldsymbol{\nu}$$

and the autocovariances are

$$\Gamma_Y(h) = \sum_{i=0}^{\infty} \mathbf{A}^{h+i} \Sigma_U (\mathbf{A}^i)',$$

where $\Sigma_U := E(U_t U_t')$. Using the $(K \times Kp)$ matrix

$$J := [I_K : 0 : \dots : 0],$$

Note que se puede obtener el proceso K -dimensional $\{\underline{Y}_t\}$ haciendo $\underline{Y}_t = J \underline{Y}_t$, y también $E[\underline{Y}_t] = J \boldsymbol{\mu}$ y $\Gamma_{\underline{Y}} = J \Gamma_Y J'$.

Finalmente, se puede verificar que

$$\det(I_{Kp} - \mathbf{A}z) = \det(I_K - A_1z - \cdots - A_pz^p)$$

y así, la condición de estabilidad del proceso $VAR(p)$ queda establecida como

$$\det(I_K - A_1z - \cdots - A_pz^p) \neq 0, \quad \text{para } |z| \leq 1,$$

y así, la forma $MA(\infty)$ del proceso $VAR(p)$ queda de la siguiente forma:

$$\underline{Y}_t = J\underline{Y}_t = J\tilde{\mu} + J \sum_{i=0}^{\infty} \mathbf{A}^i \underline{U}_{t-i},$$

y dado que el proceso \underline{U}_t satisface $\underline{U}_t = J'J\underline{U}_t$ y $J\underline{U}_t = \underline{u}_t$, tenemos que

$$J \sum_{i=0}^{\infty} \mathbf{A}^i \underline{U}_{t-i} = \sum_{i=0}^{\infty} J\mathbf{A}^i J' J \underline{U}_{t-i}$$

y escribiendo $\Phi_i = J\mathbf{A}^i J'$, tenemos

$$\sum_{i=0}^{\infty} J\mathbf{A}^i J' J \underline{U}_{t-i} = \sum_{i=0}^{\infty} \Phi_i \underline{u}_{t-i}. \quad (11.3)$$

Nota 11.6. Note que si asumimos $\{\underline{u}_t\} \sim N(0, \Sigma_u)$ un proceso ruido blanco Gaussiano, entonces el proceso $\{\underline{Y}_t\}$ también es Gaussiano.

Ejercicio 11.7. Considere el proceso $VAR(1)$ $\{\underline{Y}_t\}$ tridimensional

$$\underline{Y}_t = \underline{\nu} + \begin{bmatrix} 0.5 & 0 & 0 \\ 0.1 & 0.1 & 0.3 \\ 0 & 0.2 & 0.3 \end{bmatrix} \underline{Y}_{t-1} + \underline{u}_t.$$

Veamos si el proceso definido así, es estable. Es decir, vamos a calcular las raíces del polinomio característico:

$$\det \left(\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} - \begin{bmatrix} 0.5 & 0 & 0 \\ 0.1 & 0.1 & 0.3 \\ 0 & 0.2 & 0.3 \end{bmatrix} z \right).$$

Note que haciendo operaciones, tenemos

$$\det \left(\begin{bmatrix} 1 - 0.5z & 0 & 0 \\ -0.1z & 1 - 0.1z & -0.3z \\ 0 & -0.2z & 1 - 0.3z \end{bmatrix} \right) = (1 - 0.5z)(1 - 0.4z - 0.03z^2).$$

Note que las raíces del ese polinomio son $z_1 = 2, z_2 = 2.1525, z_3 = -15.4858$, las cuales están por fuera del círculo unitario.

Veamos como quedan las ecuaciones para cada componente:

$$Y_{1,t} = \nu_1 + 0.5Y_{1,t-1} + u_{1,t}$$

$$Y_{2,t} = \nu_2 + 0.1Y_{1,t-1} + 0.1Y_{2,t-1} + 0.3Y_{3,t-1} + u_{3,t}$$

$$Y_{3,t} = \nu_3 + 0.2Y_{2,t-1} + 0.3Y_{3,t-1} + u_{3,t}$$

Ejercicio 11.8. Considere el proceso VAR(2) $\{\underline{Y}_t\}$ bidimensional

$$\underline{Y}_t = \underline{\nu} + \begin{bmatrix} 0.5 & 0.1 \\ 0.4 & 0.5 \end{bmatrix} \underline{Y}_{t-1} + \begin{bmatrix} 0 & 0 \\ 0.25 & 0 \end{bmatrix} \underline{Y}_{t-2} + \underline{u}_t.$$

Es un proceso estable?

Tarea: Leer la sección 2.1.2 del libro [27], relacionado con la representación $MA(\infty)$ de un proceso VAR y en la sección 2.1.3 la representación de *Wold*.

Definición 11.9. Un proceso $\{\underline{Y}_t\}$ VAR(p) estable es estacionario en el sentido débil.

Nota 11.10. La condición de estabilidad usualmente se refiere a la condición de estacionariedad en procesos VAR. Un proceso inestable es no necesariamente no estacionario.

Nota 11.11. Si el proceso $\{\underline{Y}_t\}$ es un proceso VAR(P) estacionario, es decir, se puede escribir

$$\underline{Y}_t = \underline{\nu} + A_1\underline{Y}_{t-1} + \cdots + A_p\underline{Y}_{t-p} + \underline{u}_t, \quad t = 0, \pm 1, \pm 2, \dots,$$

se puede escribir introduciendo el operador de retardo L o B , es decir

$$\underline{Y}_t - A_1\underline{Y}_{t-1} - \cdots - A_p\underline{Y}_{t-p} = \underline{\nu} + \underline{u}_t$$

se puede escribir como

$$(I_k - A_1L - A_2L^2 - \cdots - A_pL^p)\underline{Y}_t = \underline{\nu} + \underline{u}_t$$

el cual se puede escribir en forma compacta

$$A(L)\underline{Y}_t = \underline{\nu} + \underline{u}_t$$

con $A(L) = I_k - A_1L - A_2L^2 - \cdots - A_pL^p$, es decir, el operador autorregresivo matricial. Ya que el proceso $\{\underline{Y}_t\}$ es estacionario, entonces se puede verificar que existe un operador $\Phi(L)$, tal que

$$A(L)\Phi(L) = I_k,$$

es decir el operador $\Phi(L)$ es el inverso de operador $A(L)$, y el cual se puede escribir como

$$\Phi(L) = A(L)^{-1} = \sum_{i=0}^{\infty} \Phi_i L^i,$$

con lo cual

$$\underline{Y}_t = \left(\sum_{i=0}^{\infty} \Phi_i L^i \right) \underline{\nu} + \sum_{i=0}^{\infty} \Phi_i \underline{u}_{t-i}.$$

Así, vale la pena decir que el operador $A(L)$ tiene inverso si $|A(z)| \neq 0$ para $|z| \leq 1$. Note que la

11.1.2 Cómputo de la la función de autocovarianza y de autocorrelación matricial de un proceso VAR estable

Note que las autocovarianzas de un proceso VAR(1) se pueden obtener en términos de su forma $MA(\infty)$. Sin embargo, en la práctica esto no es atractivo porque involucra una suma infinita. Veamos una forma recurrente de obtener estas autocovarianzas.

Autocovarianzas de un proceso VAR(1)

Vamos a considerar en proceso $\{\underline{Y}_t\}$ estable definido como

$$\underline{Y}_t = \underline{\nu} + A_1 \underline{Y}_{t-1} + \underline{u}_t$$

donde $\{\underline{u}_t\} \sim RB(Q, \Sigma_u)$.

Veamos ahora el proceso anterior ajustado por la media, es decir

$$\underline{Y}_t - \underline{\mu} = A_1(\underline{Y}_{t-1} - \underline{\mu}) + \underline{u}_t \tag{11.4}$$

con $E[\underline{Y}_t] = \underline{\mu} = (I_k - A_1)^{-1}\underline{\nu}$. Note que ahora, posmultiplicando 11.4 por $(\underline{Y}_{t-h} - \underline{\mu})'$ y tomando valor esperado nos da

$$E[(\underline{Y}_t - \underline{\mu})(\underline{Y}_{t-h} - \underline{\mu})'] = A_1 E[(\underline{Y}_{t-1} - \underline{\mu})(\underline{Y}_{t-h} - \underline{\mu})'] + E[\underline{u}_t(\underline{Y}_{t-h} - \underline{\mu})']$$

entonces tenemos las llamadas ecuaciones de Yule-Walker

Thus, for $h = 0$,

$$\Gamma_y(0) = A_1 \Gamma_y(-1) + \Sigma_u = A_1 \Gamma_y(1)' + \Sigma_u$$

and for $h > 0$,

$$\Gamma_y(h) = A_1 \Gamma_y(h - 1).$$

Figura 11.1:

Por supuesto, si $\gamma_Y(0)$ y A_1 son conocidas, entonces se puede encontrar $\Gamma_Y(h)$ para cada h de forma recurrente. De lo contrario se debe sumir que A_1 y Σ_u son conocida y encontrar a $\gamma_Y(0)$, como sigue:

If A_1 and Σ_u are given, $\Gamma_y(0)$ can be determined as follows. For $h = 1$, we get from (2.1.31), $\Gamma_y(1) = A_1\Gamma_y(0)$. Substituting $A_1\Gamma_y(0)$ for $\Gamma_y(1)$ in (2.1.30) gives

$$\Gamma_y(0) = A_1\Gamma_y(0)A_1' + \Sigma_u$$

or

$$\begin{aligned} \text{vec } \Gamma_y(0) &= \text{vec}(A_1\Gamma_y(0)A_1') + \text{vec } \Sigma_u \\ &= (A_1 \otimes A_1) \text{vec } \Gamma_y(0) + \text{vec } \Sigma_u. \end{aligned}$$

(For the definition of the Kronecker product \otimes , the vec operator and the rules used here, see Appendix A). Hence,

$$\text{vec } \Gamma_y(0) = (I_{K^2} - A_1 \otimes A_1)^{-1} \text{vec } \Sigma_u. \quad (2.1.32)$$

Note that the invertibility of $I_{K^2} - A_1 \otimes A_1$ follows from the stability of y_t because the eigenvalues of $A_1 \otimes A_1$ are the products of the eigenvalues of A_1 (see Appendix A). Hence, the eigenvalues of $A_1 \otimes A_1$ have modulus less than 1. Consequently, $\det(I_{K^2} - A_1 \otimes A_1) \neq 0$ (see Appendix A.9.1).

Figura 11.2:

Ejemplo 11.12. Consideremos el proceso VAR(1) bivariado $\{\underline{Y}_t\}$, tal que

$$\begin{bmatrix} Y_{1,t} \\ Y_{2,t} \end{bmatrix} = \begin{bmatrix} 5 \\ 3 \end{bmatrix} + \begin{bmatrix} 0.2 & 0.3 \\ -0.6 & 1.1 \end{bmatrix} \begin{bmatrix} Y_{1,t-1} \\ Y_{2,t-1} \end{bmatrix} + \begin{bmatrix} u_{1,t-1} \\ u_{2,t-1} \end{bmatrix}.$$

$$\text{y } \Sigma_u = \begin{bmatrix} 1.8 & 0.8 \\ 0.8 & 2.0 \end{bmatrix}.$$

Se puede verificar que este proceso es estacionario ya que los valores propios de la matriz $\begin{bmatrix} 0.2 & 0.3 \\ -0.6 & 1.1 \end{bmatrix}$ son 0.5 y 0.8, los cuales están por dentro del círculo unitario.

Note que un procedimiento análogo puede ser usado para encontrar las autocovarianzas de un proceso VAR(p) como sigue

$$\begin{aligned} \Gamma_y(0) &= A_1\Gamma_y(-1) + \cdots + A_p\Gamma_y(-p) + \Sigma_u \\ &= A_1\Gamma_y(1)' + \cdots + A_p\Gamma_y(p)' + \Sigma_u, \end{aligned}$$

and for $h > 0$,

$$\Gamma_y(h) = A_1\Gamma_y(h-1) + \cdots + A_p\Gamma_y(h-p).$$

Figura 11.3:

donde las condiciones iniciales se encuentran escribiendo el VAR(p) como un VAR(1) de mas grande dimensionalidad y aplicando el mismo procedimiento anterior. Para detalles, ver libro [27] Páginas 28 y 29 y detallar el ejemplo del VAR(2). Se puede verificar que la función de autocovarianza matricial es semidefinida poistiva, en el siguiente sentido:

$$\begin{aligned} & \sum_{j=0}^n \sum_{i=0}^n a'_j \Gamma_y(i-j) a_i \\ &= (a'_0, \dots, a'_n) \begin{bmatrix} \Gamma_y(0) & \Gamma_y(1) & \dots & \Gamma_y(n) \\ \Gamma_y(-1) & \Gamma_y(0) & \dots & \Gamma_y(n-1) \\ \vdots & \vdots & \ddots & \vdots \\ \Gamma_y(-n) & \Gamma_y(-n+1) & \dots & \Gamma_y(0) \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ \vdots \\ a_n \end{bmatrix} \geq 0 \end{aligned} \quad (2)$$

Figura 11.4:

Autocorrelaciones de un Proceso VAR(p) estacionario

Recordemos que las autocorrelaciones se pueden obtener a través de las autocovarianzas como sigue:

$$R_Y(h) = D^{-1} \Gamma_Y(h) D^{-1}$$

donde

$$D^{-1} = \begin{bmatrix} 1/\sqrt{\gamma_{11}(0)} & \cdots & 0 \\ & \ddots & \\ 0 & \cdots & 1/\sqrt{\gamma_{KK}(0)} \end{bmatrix}$$

donde la entrada $\rho_{ij}(h)$ de la matriz $R_Y(h)$, es la correlación entre $Y_{i,t}$ y $Y_{j,t-h}$.

Pronóstico

Vamos a suponer que el verdadero modelo generador de los datos es conocido que un conjunto de información Ω_t , contiene la información disponible hasta el tiempo t . El proceso generador de datos puede ser un VAR(p), mientras que el conjunto de información disponible puede escribirse mas específicamente como sigue:

$$\Omega_t = \{\mathcal{Y}_s | s \leq t\},$$

donde $\underline{Y}_s = (Y_{1s}, \dots, Y_{Ks})'$. El tiempo t se conoce como el origen de pronóstico, mientras que el número de periodos en el futuro para el cual se desea el pronóstico se conoce como horizonte de pronóstico.

Función de Pérdida

Por supuesto, para encontrar un pronóstico óptimo, se debe asociar con una función de costo, costo esperado o pérdida la cual debería ser minimizada. En modelos VAR, los predictores que minimizan el error cuadrático medio de pronóstico(MSE) son los más usados. Hay argumentos en favor de usar el MSE como función de pérdida, ya que los pronósticos con MSE mínimo también minimizan otras funciones de pérdida. También, para muchas funciones de pérdida los predictores óptimos son funciones simples de los predictores con mínimo MSE.

Pronósticos Puntuales

Supongamos que $\{\underline{Y}_t\}$ es un proceso VAR(p) estable. Entonces, el predictor que minimiza el MSE para el pronóstico en el horizonte h , es de hecho la esperanza condicional

$$E_t[\underline{Y}_{t+h}] = E[\underline{Y}_{t+h}|\Omega_t] = E[\underline{Y}_{t+h}|\{\underline{Y}_t | s \leq t\}]. \quad (11.5)$$

Nota 11.13. *Cuando se habla que minimiza el MSE, se refiere a que para cualquier otro predictor h -pasos adelante, digamos $\bar{\underline{Y}}_t(h)$ en el origen t , se tiene que*

$$MSE(\bar{\underline{Y}}_t(h)) \geq MSE(E_t[\underline{Y}_{t+h}]).$$

El signo de desigualdad \geq se entiende como el semi-orden de Lowener entre dos matrices simétricas ver libro [16], es decir, la diferencia entre $MSE(\bar{\underline{Y}}_t(h)) - MSE(E_t[\underline{Y}_{t+h}])$ es un matriz semidefinida positiva, Ver [27] página 34 para la deducción.

Adicionalmente, el predictor $E_t[\underline{Y}_{t+h}]$ minimiza el MSE de cada componente de \underline{Y}_{t+h} .

Para el caso del modelo VAR(p), de las propiedades de la esperanza condicional se puede demostrar que

$$E_t[\underline{Y}_{t+h}] = \nu + A_1 E_t[\underline{Y}_{t+h-1}] + \cdots + A_p E_t[\underline{Y}_{t+h-p}] \quad (11.6)$$

es el predictor óptimo de un proceso VAR(p), asumiendo que \underline{u}_t es un ruido blanco independiente para que $E_t[\underline{u}_{t+h}] = 0$.

Note que de la ecuación 11.5, se puede obtener la predicciones h -pasos adelante como sigue:

$$\begin{aligned} E_t[\underline{Y}_{t+1}] &= \underline{\nu} + A_1 \underline{Y}_t + \cdots + A_p \underline{Y}_{t-p+1} \\ E_t[\underline{Y}_{t+2}] &= \underline{\nu} + A_1 E_t[\underline{Y}_{t+1}] + A_2 \underline{Y}_t + \cdots + A_p \underline{Y}_{t-p+2} \end{aligned}$$

Nota 11.14. Note que para un proceso VAR(1), tenemos que la predicción h -pasos adelante está dado por:

$$E_t[\underline{Y}_{t+h}] = (I_k + A_1 + \cdots + A_1^{h-1})\underline{\nu} + A_1^h \underline{Y}_t.$$

Ejemplo

Assuming $y_t = (-6, 3, 5)'$ and $\nu = (0, 2, 1)'$, the following forecasts are obtained for the VAR(1) example process (2.1.14):

$$E_t(y_{t+1}) = \begin{bmatrix} 0 \\ 2 \\ 1 \end{bmatrix} + \begin{bmatrix} .5 & 0 & 0 \\ .1 & .1 & .3 \\ 0 & .2 & .3 \end{bmatrix} \begin{bmatrix} -6 \\ 3 \\ 5 \end{bmatrix} = \begin{bmatrix} -3.0 \\ 3.2 \\ 3.1 \end{bmatrix}, \quad (2.2.4a)$$

$$E_t(y_{t+2}) = (I_3 + A_1)\nu + A_1^2 y_t = \begin{bmatrix} -1.50 \\ 2.95 \\ 2.57 \end{bmatrix}, \quad (2.2.4b)$$

Tarea: Considere el proceso VAR(2) del ejemplo 11.8 con $\underline{\nu} = (0.02, 0.03)', \underline{Y}_t = (0.06, 0.03)'$ y $\underline{Y}_{t-1} = (0.055, 0.03)'$. Calcule los pronósticos 1, 2, 3 pasos adelante.

Nota 11.15. El predictor óptimo como esperanza condicional tiene las siguientes propiedades:

✓ Es un predictor insesgado, es decir, $E[\underline{Y}_{t+h} - E_t[\underline{Y}_{t+h}]] = 0$.

✓ Si $\{\underline{u}_t\}$ es ruido blanco independiente,

$$MSE(E_t[\underline{Y}_{t+h}]) = MSE[E_t[\underline{Y}_{t+h}]|\underline{Y}_t, \underline{Y}_{t-1}, \dots],$$

es decir, el MSE del el predictor iguala el MSE condicional dado $\underline{Y}_t, \underline{Y}_{t-1}, \dots$.

Si el proceso de ruido $\{\underline{u}_t\}$ no es independiente, entonces $E_t[\underline{u}_{t+h}]$ en general no es igual a cero y así la fórmula 11.6 no es válida ya que depende de este supuesto.

El predictor lineal con MSE mínimo

Si el supuesto sobre $\{\underline{u}_t\}$ no es ruido blanco independiente, entonces se requiere de supuestos adicionales para encontrar el mejor(óptimo) predictor de un proceso VAR(p). Sin esto se debe considerar un predictor óptimo entre la familia de funciones lineales de $\underline{Y}_t, \underline{Y}_{t-1}, \dots$. Si el proceso es un VAR(1), se puede verificar que el mejor(tal que minimiza el el MSE) predictor lineal $\underline{Y}_t(h)$ con base en $\underline{Y}_t, \underline{Y}_{t-1}, \dots$ es

$$\underline{Y}_t(h) = A_1^h \underline{Y}_t = A_1 \underline{Y}_t(h-1).$$

y el error de pronóstico es $\sum_{i=0}^{h-1} A_1^i \underline{u}_{t+h-i}$, con la matriz de error cuadrático medio dada por

$$\Sigma_{\underline{Y}}(h) = MSE(\underline{Y}_t(h)) = \sum_{i=0}^{h-1} A_1^i \Sigma_{\underline{u}}(A_1^i)' = MSE[\underline{Y}_t(h-1)] + A_1^{h-1} \Sigma_{\underline{u}}(A_1^{h-1})'.$$

Tarea: Cómo quedan las expresiones del mejor predictor lineal para un VAR(p)? ver [27] Página 36.

Como en el caso del VAR(1), $\underline{Y}_t(h) = \mathbf{A}^h \underline{Y}_t = \mathbf{A} \underline{Y}_t(h-1)$ y el cual tiene forma específica

$$\begin{bmatrix} \underline{Y}_t(h) \\ \underline{Y}_t(h-1) \\ \vdots \\ \underline{Y}_t(h-p+1). \end{bmatrix}$$

Note que el predictor óptimo del proceso $\{\underline{Y}_t\}$ en el origen t se puede obtener de forma recursiva como sigue:

$$\underline{Y}_t = J \mathbf{A} \underline{Y}_t(h-1) = [A_1, \dots, A_p] \underline{Y}_t(h-1) = A_1 \underline{Y}_t(h-1) + \dots + A_p \underline{Y}_t(h-p).$$

Note que el predictor óptimo se puede escribir en términos de la representación $MA(\infty)$ como

$$\underline{Y}_t(h) = \underline{\mu} + \sum_{i=0}^{\infty} \Phi_{h+i} \underline{u}_{t-i}, \quad (11.7)$$

y junto con 11.3, tenemos que la covarianza del error de pronóstico o la matriz MSE se obtiene como sigue:

$$\Sigma_{\underline{Y}}(h) = MSE[\underline{Y}_t(h)] = \sum_{i=0}^{h-1} \Phi_i \Sigma_{\underline{u}} \Phi_i' = \Sigma_{\underline{Y}}(h-1) + \Phi_{h-1} \Sigma_{\underline{u}} \Phi_{h-1}'.$$

Nota 11.16. Con la fórmula anterior es evidente que el MSE es monótonamente no decreciente, y cuando $h \rightarrow \infty$ se puede verificar que

$$\Sigma_{\underline{Y}}(h) \xrightarrow[h \rightarrow \infty]{} \Sigma_{\underline{Y}}.$$

Ejemplo 11.17. Consideremos el procesos

$$\underline{Y}_t = \underline{\nu} + \begin{bmatrix} 0.5 & 0 & 0 \\ 0.1 & 0.1 & 0.3 \\ 0 & 0.2 & 0.3 \end{bmatrix} \underline{Y}_{t-1} + \underline{u}_t.$$

con

$$\Sigma_{\underline{u}} = \begin{bmatrix} 2.25 & 0 & 0 \\ 0.0 & 1.0 & 0.5 \\ 0 & 0.5 & 0.74 \end{bmatrix}$$

, computar los errores de predicción 1,2 y 3 pasos adelante.

Intervalos de Pronóstico y Regiones de Pronóstico

Inicialmente supongamos que el proceso de ruido $\{\underline{u}_t\}$ es Gaussiano, y por lo tanto el proceso $\{\underline{Y}_t\}$ también lo es. Bajo este supuesto, el error de pronóstico h -pasos adelante tiene la siguiente estructura(para el caso del VAR(1)):

$$\underline{Y}_t - \underline{Y}_t(h) = \sum_{i=0}^{h-1} A_1^i \underline{u}_{t+h-i} \sim N(\underline{0}, \Sigma_{\underline{Y}}(h))$$

mientras que para el caso del proceso VAR(p), tiene una forma análoga

$$\underline{Y}_t - \underline{Y}_t(h) = J[\underline{Y}_t - \underline{Y}_t(h)] = J \left[\sum_{i=0}^{h-1} A^i \underline{U}_{t+h-i} \right] \sim N(\underline{0}, \Sigma_{\underline{Y}}(h)).$$

Con lo cual, para cada componente individual k del vector de errores se tiene que

$$\frac{Y_{k,t+h} - Y_{k,t}(h)}{\sigma_k(h)} \sim (0, 1),$$

donde $Y_{k,t}(h)$ es la k -ésima componente del vector $\underline{Y}_t(h)$ y $\sigma_k(h)$ es la raíz cuadrada del k -ésimo elemento de la diagonal de la matriz $\Sigma_{\underline{Y}}(h)$. Por lo tanto un intervalo de pronóstico individual h -pasos adelante del $100(1-\alpha)\%$, está dado por

$$Y_{k,t}(h) \pm Z_{1-\alpha/2} \sigma_k(h).$$

Tarea 11.18. *Cómo se construyen regiones de pronóstico?*

11.1.3 Análisis Estructural con Modelos VAR

Dado que un modelo VAR representa correlaciones entre un conjunto de variables, ellos son usados para analizar ciertos aspectos de las relaciones entre las variables de interés.

Causalidad de Granger

La definición de Causalidad dada por Granger en el año de 1969 ha tenido un gran impacto en los últimos años, y en especial, dado que es fácil de trabajar para modelos VAR, se ha vuelto bastante popular. La idea es que una causa no puede venir después del efecto. Así, si una variable x afecta a la variable z , esta primera debería mejorar las predicciones de la última variable.

Para desarrollar la idea, consideremos que Ω_t es el conjunto de información que contiene toda la información relevante disponible en el universo hasta el periodo de tiempo t . Consideremos ahora que $\underline{Z}_t(h|\Omega_t)$ es el predictor óptimo(en el sentido del MSE) h -pasos adelante del proceso $\{\underline{Z}_t\}$ en el origen t , basado en la información Ω_t . En ese sentido, el MSE del pronóstico correspondiente a este predictor se denota por $\Sigma_{\underline{Z}}(h|\Omega_t)$. Establecido lo anterior, tenemos la siguiente definición:

Definición 11.19. *Diremos que el proceso $\{\underline{X}_t\}$ causa a $\{\underline{Z}_t\}$ en el sentido de Granger si*

$$\Sigma_{\underline{Z}}(h|\Omega_t) \leq \Sigma_{\underline{Z}}(h|\Omega_t - \{\underline{X}_s | s \leq t\}) \quad \text{para al menos un } h = 1, 2, \dots$$

Nota 11.20. *Que $\{\underline{X}_t\}$ cause a $\{\underline{Z}_t\}$ en el sentido de Granger, quiere decir que \underline{Z}_t puede ser predicha mas eficientemente si la información de $\{\underline{X}_t\}$ es tenida en cuenta en adición a toda la otra información en el universo, por lo tanto también se usar decir que $\{\underline{X}_t\}$ es Granger-causal para $\{\underline{Z}_t\}$.*

Nota 11.21. *Si $\{\underline{X}_t\}$ cause a $\{\underline{Z}_t\}$ en el sentido de Granger, y también $\{\underline{Z}_t\}$ causa a $\{\underline{X}_t\}$ en el sentido de Granger entonces al proceso $\{(\underline{X}'_t, \underline{Z}'_t)'\}$ se le conoce como un sistema retroalimentado.*

Definición 11.22. *Diremos que existe causalidad instantánea entre $\{\underline{Z}_t\}$ y $\{\underline{X}_t\}$ si*

$$\Sigma_{\underline{Z}}(1|\Omega_t \cup \{\underline{X}_{t+1}\}) \neq \Sigma_{\underline{Z}}(1|\Omega_t),$$

es decir, en el periodo t , al añadir \underline{X}_{t+1} al conjunto de información este ayuda a mejorar el pronóstico de \underline{Z}_{t+1} .

Nota 11.23. ✓ Note que la definición de causalidad está relacionada directamente con el MSE, sin embargo puede utilizarse una medida diferente de optimalidad, lo cual hace que la definición puede cambiar.

- ✓ Desde el punto de vista práctico, el conjunto de información Ω_t debe definirse de otra manera ya que toda la información relevante en el universo no está disponible. Por lo tanto el conjunto de información se re-define como $\Omega_t = \{\underline{Z}_s, \underline{X}_s | s \leq t\}$.
- ✓ En la práctica también se considera el mejor(óptimo) predictor lineal en vez del predictor óptimo, es decir, $\underline{Z}_t(h|\Omega)$ es reemplazado por el predictor h -pasos adelante que minimiza el MSE basado en la información $\{\underline{Z}_s, \underline{X}_s | s \leq t\}$, mientras que $\underline{Z}_t(h|\Omega - \{\underline{X}_s | s \leq t\})$ es reemplazado por el predictor lineal h -pasos adelante que minimiza el MSE basado en $\{\underline{Z}_s | s \leq t\}$.

Caracterización de la Causalidad de Granger en modelos VAR

La idea inicial va a consistir en como obtenemos los pronósticos necesarios, y así como el MSE para ser comparados y caracterizar la causalidad. Vamos a suponer que el proceso K -dimensional $\{\underline{Y}_t\}$ tiene representación canónica o $MA(\infty)$, es decir:

$$\underline{Y}_t = \underline{\mu} + \sum_{i=0}^{\infty} \Phi_i \underline{u}_{t-i} = \underline{\mu} + \Phi(L) \underline{u}_t, \quad \Phi_0 = I_K \quad (11.8)$$

donde $\{\underline{u}_t\} \sim RB(\underline{0}, \Sigma_{\underline{u}})$, y la matriz $\Sigma_{\underline{u}}$ es no singular.

Vamos a considerar un partición del proceso $\{\underline{Y}_t\} = (\underline{Z}_t, \underline{X}_t)$, cuyas dimensiones de los subprocesos $\{\underline{Z}_t\}$ y \underline{X}_t son M y $M-K$ respectivamente. La representación 11.8 para la partición queda como:

$$\underline{Y}_t = \begin{bmatrix} \underline{Z}_t \\ \underline{X}_t \end{bmatrix} = \begin{bmatrix} \underline{\mu}_1 \\ \underline{\mu}_2 \end{bmatrix} + \begin{bmatrix} \Phi_{11}(L) & \Phi_{12}(L) \\ \Phi_{21}(L) & \Phi_{22}(L) \end{bmatrix} \begin{bmatrix} \underline{u}_{1t} \\ \underline{u}_{2t} \end{bmatrix}. \quad (11.9)$$

el predictor un paso adelante de \underline{Z}_t basado en \underline{Y}_t se puede obtener de \underline{Y}_t como

$$\underline{Z}_t(1 | \{\underline{Y}_s | s \leq t\}) = [I_M : 0] \underline{Y}_t(1) \quad (11.10)$$

$$= \underline{\mu}_1 + \sum_{i=1}^{\infty} \Phi_{11,i} \underline{u}_{1,t+1-i} + \sum_{i=1}^{\infty} \Phi_{12,i} \underline{u}_{2,t+1-i} \quad (11.11)$$

Con lo cual el error de pronóstico es

$$\underline{Z}_{t+1} - \underline{Z}_t(1|\{\underline{Y}_s | s \leq t\}).$$

El anterior procedimiento permite encontrar la predicción con base en toda la información, ahora vamos a hallar la predicción con base únicamente en la información del proceso $\{\underline{Z}_t\}$. Dado que un sub-proceso de un proceso estacionario es también estacionario, entonces también tiene un representación $MA(\infty)$, digamos de la siguiente forma

$$\underline{Z}_t = \underline{\mu}_1 + \sum_{j=0}^{\infty} F_j \underline{v}_{t-j}.$$

Con esta representación se puede verificar que

$$\underline{Z}_t(1|\{\underline{Z}_s | s \leq t\}) = \underline{\mu}_1 + \sum_{i=1}^{\infty} F_i \underline{v}_{t+1-i}$$

y el error de pronóstico es

$$\underline{Z}_{t+1} - \underline{Z}_t(1|\{\underline{Z}_s | s \leq t\}) = \underline{v}_{t+1}.$$

Nota 11.24. Los dos predictores serán idénticos si y sólo si $\underline{v}_t = \underline{u}_{1,t}$ para todo t . Lo cual necesariamente implica que los dos predictores son equivalentes si el proceso $\{\underline{Z}_t\}$ tiene una representación MA igual, es decir

$$\begin{aligned} z_t &= \mu_1 + \sum_{i=0}^{\infty} F_i u_{1,t-i} = \mu_1 + \sum_{i=0}^{\infty} [\Phi_{11} : 0] u_{t-i} \\ &= \mu_1 + \sum_{i=0}^{\infty} [\Phi_{11,i} : \Phi_{12,i}] u_{t-i} \\ &= \mu_1 + \sum_{i=0}^{\infty} \Phi_{11,i} u_{1,t-i} + \sum_{i=1}^{\infty} \Phi_{12,i} u_{2,t-i}. \end{aligned}$$

Figura 11.5:

Dado que $\Phi_0 = I_K$, entonces $\Phi_{12,0} = 0$, por eso empieza en 1 la segunda sumatoria. Cómo la representación de un proceso MA es única, entonces igualando la primera igualdad con la última igualdad tenemos que esta representación es única siempre que $F_i = \Phi_{11,i}$ y $\Phi_{12,i} = 0$ para $i = 1, 2, \dots$, con lo cual tenemos el siguiente resultado:

Proposición 11.25. La caracterización de no-causalidad de Granger. Sea $\{\underline{Y}_t\}$ un proceso VAR que tiene representación canónica $MA(\infty)$ como

en 11.9 y 11.8. Entonces

$$\underline{Z}_t(1|\{\underline{Y}_s|s \leq t\}) = \underline{Z}_t(1|\{\underline{Z}_s|s \leq t\}) \Leftrightarrow \Phi_{12,i} = 0 \text{ para } i = 1, 2, \dots \quad (11.12)$$

Note que esta caracterización no sólo es válida para el un proceso VAR, sino también para cualquier proceso que tenga la forma canónica $MA(\infty)$. También, basados en 11.7 se puede verificar que la igualdad de los predictores 1-paso adelante implica la igualdad de los predictores h -pasos adelantepara $h = 2, 3, \dots$.

Nota 11.26. *Dado que estamos interesados en la causalidad de Granger para un proceso $VAR(p)$ estable,*

$$\underline{Y}_t = \begin{bmatrix} \underline{Z}_t \\ \underline{X}_t \end{bmatrix} = \begin{bmatrix} \underline{v}_1 \\ \underline{v}_2 \end{bmatrix} + \begin{bmatrix} A_{11,1} & \color{blue}{A_{12,1}} \\ A_{21,1} & A_{22,1} \end{bmatrix} \begin{bmatrix} \underline{Z}_{t-1} \\ \underline{X}_{t-1} \end{bmatrix} + \dots + \begin{bmatrix} A_{11,p} & \color{blue}{A_{12,p}} \\ A_{21,p} & A_{22,p} \end{bmatrix} \begin{bmatrix} \underline{Z}_{t-p} \\ \underline{X}_{t-p} \end{bmatrix} + \begin{bmatrix} \underline{u}_{1t} \\ \underline{u}_{2t} \end{bmatrix} \quad (11.13)$$

la condición 11.13 se satisface si y sólo si

$$A_{12,i} = 0$$

para $i = 1, 2, \dots, p$ debido a las recurrencias del $MA(\infty)$

$$\Phi_0 = I_K$$

$$\Phi_i = \sum_{j=1}^i \Phi_{i-j} A_j, \quad i = 1, 2, \dots$$

o porque el inverso del operador $\begin{bmatrix} \Phi_{11}(L) & 0 \\ \Phi_{21}(L) & \Phi_{22}(L) \end{bmatrix}$ es
 $\begin{bmatrix} \Phi_{11}(L)^{-1} \\ -\Phi_{22}(L)^{-1}\Phi_{21}(L)\Phi_{11}(L)^{-1} & \Phi_{22}(L)^{-1} \end{bmatrix}$.

De los anterior tenemos el siguiente resultado que caracteriza la no-causalidad en procesos $VAR(p)$:

Corolario 11.27. *Si $\{\underline{Y}_t\}$ es un proceso estable $VAR(p)$ como en 11.13 con matriz covarianza del ruido Σ_u no-singular, entonces*

$$\begin{aligned} \underline{Z}_t(1|\{\underline{Y}_s|s \leq t\}) &= \underline{Z}_t(1|\{\underline{Z}_s|s \leq t\}), \quad h = 1, 2, \dots, \\ &\Leftrightarrow A_{12,i} = 0 \text{ para } i = 1, 2, \dots, p. \end{aligned} \quad (11.14)$$

o alternativamente,

$$\begin{aligned} \underline{X}_t(h|\{\underline{Y}_s|s \leq t\}) &= \underline{X}_t(h|\{\underline{X}_s|s \leq t\}), \quad h = 1, 2, \dots, \\ &\Leftrightarrow A_{21,i} = 0 \text{ para } i = 1, 2, \dots, p. \end{aligned} \quad (11.15)$$

Ejemplo 11.28. Considere el siguiente proceso VAR(1) trideimensional

$$\begin{bmatrix} Y_{1,t} \\ Y_{2,t} \\ Y_{3,t} \end{bmatrix} = \underline{\nu} + \begin{bmatrix} 0.5 & 0 & 0 \\ 0.1 & 0.1 & 0.3 \\ 0 & 0.2 & 0.3 \end{bmatrix} \begin{bmatrix} Y_{1,t-1} \\ Y_{2,t-1} \\ Y_{3,t-1} \end{bmatrix} + \underline{u}_t.$$

si se considera que $\underline{X}_t = (Y_{2,t}, Y_{3,t})'$ y $Z_t = Y_{1,t}$, entonces \underline{X}_t no causa en el sentido de Granger a Z_t ya que $A_{12,1} = \mathbf{0}$. Por otro lado, Z_t causa en el sentido de Granger a \underline{X}_t ya que $A_{21,1} = (0.1, 0)'$, es decir la partición para la configuración de la matriz de coeficientes que permite chequear esta causalidad es como sigue

$$\left[\begin{array}{c|cc} 0.5 & 0 & 0 \\ - & - & - \\ 0.1 & 0.1 & 0.3 \\ 0 & 0.2 & 0.3 \end{array} \right].$$

Si asumimos que $Y_{1,t}$ es la inversión, $Y_{2,t}$ es el ingreso y $Y_{3,t}$ es el consumo y agrupamos las variables como antes, tenemos que el ingreso y el consumo con causan a la inversión, pero la inversión si causa ingreso/consumo. Si ahora agrupamos así:

$\underline{Z}_t = (Y_{1,t}, Y_{2,t})'$ y $X_t = Y_{3,t}$, entonces la partición de la matriz queda

$$\left[\begin{array}{cc|c} 0.5 & 0 & 0 \\ 0.1 & 0.1 & 0.3 \\ - & - & - \\ 0 & 0.2 & 0.3 \end{array} \right],$$

con lo cual el consumo causa a la inversión/ingreso y vice-versa.

Nota 11.29. Dado que se ha caracterizado la causalidad de Granger para dos grupos de variables, entonces no se puede hablar de causalidad entre dos variables en un sistema de tres dimensiones.

Ejemplo 11.30. Considere el proceso VAR(2)

$$\begin{bmatrix} y_{1,t} \\ y_{2,t} \end{bmatrix} = \nu + \begin{bmatrix} .5 & .1 \\ .4 & .5 \end{bmatrix} \begin{bmatrix} y_{1,t-1} \\ y_{2,t-1} \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ .25 & 0 \end{bmatrix} \begin{bmatrix} y_{1,t-2} \\ y_{2,t-2} \end{bmatrix} + u_t,$$

donde $Y_{1,t}$ es la tasa de la inflación y $Y_{2,t}$ es la tasa de interés. Cuál variable causa a cuál?

Proposición 11.31. *Caracterización de la causalidad instantánea.*

Consideré que el proceso $\{\underline{Y}_t\}$ es como en 11.9 con matriz de covarianza de las innovaciones $\Sigma_{\underline{u}}$ no singular, o como la representación MA ortogonal

$$\underline{Y}_t = \underline{\mu} + \sum_{i=0}^{\infty} \Phi_i P P^{-1} \underline{u}_{t-i} = \underline{\mu} + \sum_{i=0}^{\infty} \Theta_i \underline{w}_{t-i} \quad (11.16)$$

con $\Sigma_{\underline{u}} = P P'$ y $\Theta_i = \Phi_i P$, mientras que $\underline{w}_t = P^{-1} \underline{u}_t$, tal que $\Sigma_{\underline{w}} = I_k$ y P es una matriz triangular inferior (es decir, tiene descomposición de Cholesky). Entonces no existe causalidad instantánea entre Z_t y X_t si y sólo si

$$E(\underline{u}_{1t} \underline{u}'_{2t}) = \mathbf{0}, \quad (11.17)$$

es decir \underline{u}_{1t} y \underline{u}_{2t} son no correlacionados.

Ejemplo 11.32. Si para el caso del proceso VAR(1) tridimensional se considera que la matriz de covarianza de los errores es

$$\begin{bmatrix} 2.25 & 0 & 0 \\ 0 & 1.0 & 0.5 \\ 0 & 0.5 & 0.74 \end{bmatrix}$$

no hay causalidad instantánea entre el ingreso/consumo y la inversión.

Nota 11.33. Note que, la falta de una relación causal de Granger entre un grupo de variables y el resto de las variables no puede interpretarse necesariamente como una falta de relación de causa y efecto.

Análisis de Respuesta al Impulso

El análisis de respuesta al impulso es otro enfoque de causalidad y consiste en ver si una variable es impulsada, como esta puede generar una respuesta en otras variables. Es decir, si hay una reacción de una variable al impulso de otra, podemos decir que la última es causal de la última. En general se estudiará este tipo de causalidad rastreando el efecto de un choque exógeno o innovación en una de las variables sobre algunas o todas las otras variables. Sin embargo, no ampliaremos este enfoque, para mas detalles ver libro [27] la sección 2.3.2.

Tarea 11.34. Del libro de [27] realizar problemas 2.3, 2.4, 2.5(a,b,c), 2.6(a-d)

11.2 Estimación de Procesos Autorregresivo Vectoriales

Vamos a considerar que tenemos una serie de tiempo K -dimensional de longitud T $\underline{Y}_1, \dots, \underline{Y}_T$, tal que fue generada por un proceso VAR(P) estacionario

$$\underline{Y}_t = \underline{\nu} + A_1 \underline{Y}_{t-1} + \dots + A_p \underline{Y}_{t-p} + \underline{u}_t \quad (11.18)$$

tal que $\underline{\nu} = (\nu_1, \dots, \nu_K)'$ el vector interceptos , las A_i son matrices de coeficientes $K \times K$, y el proceso $\{\underline{u}_t\}$ es un ruido blanco con matriz de covarianza $\Sigma_{\underline{u}}$. Por lo tanto, los parámetros del modelo VAR(p) son $\underline{\nu}, A_1, A_2, \dots, A_p, \Sigma_{\underline{u}}$. Básicamente existen tres formas básicas para el procedimiento de estimación: mínimos cuadrados multivariados, Yule-Walker y máxima verosimilitud. Nos enfocaremos del estimador de mínimos cuadrados multivariados, dado que bajo Gaussianidad es equivalente al de máxima verosimilitud.

Estimación de Mínimos Cuadrados Multivariados

Vamos a considerar que tenemos una pre-muestra de tamaño p o p condiciones iniciales $\underline{Y}_{-p+1}, \dots, \underline{Y}_0$. Se define entonces lo siguiente:

$$\begin{aligned} Y &:= (y_1, \dots, y_T) && (K \times T), \\ B &:= (\nu, A_1, \dots, A_p) && (K \times (Kp+1)), \\ Z_t &:= \begin{bmatrix} 1 \\ y_t \\ \vdots \\ y_{t-p+1} \end{bmatrix} && ((Kp+1) \times 1), \\ Z &:= (Z_0, \dots, Z_{T-1}) && ((Kp+1) \times T), \\ U &:= (u_1, \dots, u_T) && (K \times T), \\ \mathbf{y} &:= \text{vec}(Y) && (KT \times 1), \\ \boldsymbol{\beta} &:= \text{vec}(B) && ((K^2p+K) \times 1), \\ \mathbf{b} &:= \text{vec}(B') && ((K^2p+K) \times 1), \\ \mathbf{u} &:= \text{vec}(U) && ((KT \times 1)). \end{aligned}$$

Figura 11.6:

Con esta definición, claramente B o $\boldsymbol{\beta}$ son los parámetros del modelo. Note que para $t = 1, \dots, T$, el VAR(p) puede escribirse en forma compacta como

$$Y = BZ + U \quad (11.19)$$

Se puede aplicar el operador vec a ambos lados de la igualdad, lo cual se obtiene

$$\begin{aligned}\text{vec}(Y) &= \text{vec}(BZ) + \text{vec}(U) \\ &= (Z' \otimes I_K)\text{vec}(B) + \text{vec}(U)\end{aligned}$$

lo cual, con definiciones hechas anteriormente, tenemos que

$$\mathbf{y} = (Z' \otimes I_K)\boldsymbol{\beta} + \mathbf{u} \quad (11.20)$$

y así la matriz de covarianzas del vector \mathbf{u} es

$$\mathbf{u} = I_T \otimes \Sigma_{\underline{u}}. \quad (11.21)$$

La estimación de mínimos cuadrados multivariados(o GLS) de $\boldsymbol{\beta}$ se entiende que que se elige el estimador que minimice:

$$\begin{aligned}S(\boldsymbol{\beta}) &= \mathbf{u}'(I_T \otimes \Sigma_{\underline{u}})^{-1}\mathbf{u} = \mathbf{u}'(I_T \otimes \Sigma_{\underline{u}}^{-1})\mathbf{u} \\ &= [\mathbf{y} - (Z' \otimes I_K)\boldsymbol{\beta}]'(I_T \otimes \Sigma_{\underline{u}}^{-1})[\mathbf{y} - (Z' \otimes I_K)\boldsymbol{\beta}] \\ &= \text{vec}(Y - BZ)'(I_T \otimes \Sigma_{\underline{u}}^{-1})\text{vec}(Y - BZ) \\ &= \text{tr}[(Y - BZ)' \Sigma_{\underline{u}}^{-1} (Y - BZ)].\end{aligned}$$

In order to find the minimum of this function we note that

$$\begin{aligned}S(\boldsymbol{\beta}) &= \mathbf{y}'(I_T \otimes \Sigma_{\underline{u}}^{-1})\mathbf{y} + \boldsymbol{\beta}'(Z \otimes I_K)(I_T \otimes \Sigma_{\underline{u}}^{-1})(Z' \otimes I_K)\boldsymbol{\beta} \\ &\quad - 2\boldsymbol{\beta}'(Z \otimes I_K)(I_T \otimes \Sigma_{\underline{u}}^{-1})\mathbf{y} \\ &= \mathbf{y}'(I_T \otimes \Sigma_{\underline{u}}^{-1})\mathbf{y} + \boldsymbol{\beta}'(ZZ' \otimes \Sigma_{\underline{u}}^{-1})\boldsymbol{\beta} - 2\boldsymbol{\beta}'(Z \otimes \Sigma_{\underline{u}}^{-1})\mathbf{y}.\end{aligned}$$

Figura 11.7:

con esto se puede ver que la función $S(\boldsymbol{\beta})$ se puede minimizar como sigue:

$$\frac{\partial S(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = 2(ZZ' \otimes \Sigma_{\underline{u}}^{-1})\boldsymbol{\beta} - 2(Z \otimes \Sigma_{\underline{u}}^{-1})\mathbf{y}.$$

Equating to zero gives the *normal equations*

$$(ZZ' \otimes \Sigma_{\underline{u}}^{-1})\hat{\boldsymbol{\beta}} = (Z \otimes \Sigma_{\underline{u}}^{-1})\mathbf{y}$$

and, consequently, the LS estimator is

$$\begin{aligned}\hat{\boldsymbol{\beta}} &= ((ZZ')^{-1} \otimes \Sigma_{\underline{u}})(Z \otimes \Sigma_{\underline{u}}^{-1})\mathbf{y} \\ &= ((ZZ')^{-1}Z \otimes I_K)\mathbf{y}.\end{aligned}$$

The Hessian of $S(\boldsymbol{\beta})$,

$$\frac{\partial^2 S}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'} = 2(ZZ' \otimes \Sigma_{\underline{u}}^{-1}),$$

Figura 11.8:

Se puede verificar que la matriz Hessiana es definida positiva, lo cual confirma que en efecto $\hat{\boldsymbol{\beta}}$ es el vector que minimiza la suma de cuadrados.

Nota 11.35. Para que los resultados anteriores se cumplan se debe satisfacer que ZZ' sea no singular. También, note que el estimador de mínimos cuadrados multivariados $\hat{\beta}$ es idéntico al estimador de mínimos cuadrados ordinarios obtenidos de minimizar

$$\bar{S}(\beta) = \underline{u}' \underline{u} = [\underline{y} - (Z' \otimes I_k)\beta]' [\underline{y} - (Z' \otimes I_k)\beta] \quad (11.22)$$

Nota 11.36. Otra forma de escribir el estimador de mínimos cuadrados es como sigue:

$$\hat{\mathbf{b}} = \text{vec}(\hat{B}') = (I_K \otimes (ZZ')^{-1}Z)\text{vec}(Y'), \quad (11.23)$$

la cual permite verificar que el estimador de mínimos cuadrados multivariados es equivalente a la estimación de mínimos cuadrados ordinarios de cada una de las K ecuaciones del modelo VAR(p) en forma separada, es decir, para un $k = 1, \dots, K$, las ecuaciones por separado son:

$$Y_{(k)} = Z'b_k + u_{(k)}$$

con $Y_{(k)} = (Y_{k1}, \dots, Y_{kT})'$ y $\mathbf{b}' = (b'_1, \dots, b'_k)$.

Esto es lo que usualmente hacen los softwares de computadora. Finalmente una, última forma de escribir el estimador de mínimos cuadrados es:

$$\hat{B} = YZ'(ZZ')^{-1}.$$

Nota 11.37. Para que las propiedades asintóticas se cumplan, si debe satisfacer que:

$$\Gamma = \text{plim} ZZ'/T \quad \text{exista y sea no singular}$$

y

$$\frac{1}{\sqrt{T}} \sum_{t=1}^T \text{vec}(\underline{u}_t Z'_{t-1}) = \frac{1}{\sqrt{T}} \text{vec}(UZ') = \frac{1}{\sqrt{T}} (Z \otimes I_k) \underline{u} \xrightarrow[T \rightarrow \infty]{d} N(0, \Gamma \otimes \Sigma_{\underline{u}}).$$

Propiedades Asintóticas de los Estimadores de mínimos cuadrados

Proposición 11.38. Proposición acerca de la distribución asintótica de los estimadores de mínimos cuadrados.

Let y_t be a stable, K -dimensional $\text{VAR}(p)$ process as in (3.1.1) with standard white noise residuals, $\widehat{B} = YZ'(ZZ')^{-1}$ is the LS estimator of the VAR coefficients B and all symbols are as defined in (3.2.1). Then,

$$\text{plim } \widehat{B} = B$$

and

$$\sqrt{T}(\widehat{\beta} - \beta) = \sqrt{T} \text{ vec}(\widehat{B} - B) \xrightarrow{d} \mathcal{N}(0, \Gamma^{-1} \otimes \Sigma_u) \quad (3.2.15)$$

or, equivalently,

$$\sqrt{T}(\widehat{\mathbf{b}} - \mathbf{b}) = \sqrt{T} \text{ vec}(\widehat{B}' - B') \xrightarrow{d} \mathcal{N}(0, \Sigma_u \otimes \Gamma^{-1}), \quad (3.2.16)$$

where $\Gamma = \text{plim } ZZ'/T$. ■

Figura 11.9:

Ahora procedemos a mirar las propiedades de los estimadores de la matriz de covarianza $\Sigma_{\underline{u}}$, además note que es necesario estimar también a Γ . Note que un estimador para consistente para Γ es

$$\widehat{\Gamma} = ZZ'/T,$$

mientras que un estimador para $\Sigma_{\underline{u}}$ está dado por

$$\widetilde{\Sigma}_{\underline{u}} = \frac{1}{T}Y(I_T - Z'(ZZ')^{-1}Z)Y' = \frac{1}{T}(Y - \widehat{B}Z)(Y - \widehat{B}Z)'.$$

Usualmente un ajuste por grados de libertad es requerido, por lo tanto un estimador queda de la siguiente forma:

$$\hat{\Sigma}_{\underline{u}} = \frac{T}{T - Kp - 1}\widetilde{\Sigma}_{\underline{u}},$$

sin embargo, los dos estimadores son asintóticamente equivalentes, esto es para el caso de muestras finitas.

Proposición 11.39. *Proposición acerca de las propiedades asintóticas de la matriz de covarianza de los errores.*

Let y_t be a stable, K -dimensional $\text{VAR}(p)$ process as in (3.1.1) with standard white noise innovations and let \bar{B} be an estimator of the VAR coefficients B so that $\sqrt{T} \text{vec}(\bar{B} - B)$ converges in distribution. Furthermore, using the symbols from (3.2.1), suppose that

$$\bar{\Sigma}_u = (Y - \bar{B}Z)(Y - \bar{B}Z)'/(T - c),$$

where c is a fixed constant. Then

$$\text{plim} \sqrt{T}(\bar{\Sigma}_u - UU'/T) = 0. \quad (3.2.20)$$

■

Figura 11.10:

Nota 11.40. *Se puede verificar que las distribuciones límite de los estimadores de mínimos cuadrados para B es independiente de la distribución límite del estimadores matriz de covarianza de los errores.*

se puede verificar que los estimadores $\tilde{\Sigma}_{\underline{u}}$ y $\hat{\Sigma}_{\underline{u}}$ son asintóticamente equivalentes y consistentes para $\Sigma_{\underline{u}}$.

Nota 11.41. *Note que si el proceso $\{\underline{Y}_t\}$ con ruido blanco estándar, entonces se obtiene directamente que*

$$\frac{(\hat{\beta}_i - \beta_i)}{\hat{s}_i} \sim N(0, 1)$$

de forma asintótica. Donde $\hat{\beta}_i$ y β_i son las i -ésimas componentes de $\hat{\beta}$ y β , mientras que \hat{s}_i es la raíz cuadrada del i -ésimo elemento de la diagonal de

$$(ZZ')^{-1} \otimes \hat{\Sigma}_{\underline{u}}.$$

11.3 Estimadores de Máxima Verosimilitud

Vamos a asumir que el proceso $\{\underline{Y}_t\}$ es Gaussiano. Con esto tenemos lo siguiente:

$$\mathbf{u} = \text{vec}(U) = \begin{bmatrix} u_1 \\ \vdots \\ u_T \end{bmatrix} \sim \mathcal{N}(0, I_T \otimes \Sigma_u). \quad (3.4.1)$$

In other words, the probability density of \mathbf{u} is

$$f_{\mathbf{u}}(\mathbf{u}) = \frac{1}{(2\pi)^{KT/2}} |I_T \otimes \Sigma_u|^{-1/2} \exp \left[-\frac{1}{2} \mathbf{u}' (I_T \otimes \Sigma_u^{-1}) \mathbf{u} \right]. \quad (3.4.2)$$

Moreover,

$$\begin{aligned} \mathbf{u} &= \begin{bmatrix} I_K & 0 & \dots & 0 & \dots & \dots & 0 \\ -A_1 & I_K & & 0 & \dots & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots & & & \vdots \\ -A_p & -A_{p-1} & \dots & I_K & & & 0 \\ 0 & -A_p & & & \ddots & & \vdots \\ \vdots & & \ddots & & \ddots & & \vdots \\ 0 & 0 & \dots & -A_p & \dots & \dots & I_K \end{bmatrix} (\mathbf{y} - \boldsymbol{\mu}^*) \\ &+ \begin{bmatrix} -A_1 & -A_2 & \dots & -A_p \\ -A_2 & -A_3 & \dots & 0 \\ \vdots & & & \vdots \\ -A_p & 0 & \dots & 0 \\ \vdots & & & \vdots \\ 0 & 0 & \dots & 0 \end{bmatrix} (Y_0 - \boldsymbol{\mu}), \end{aligned} \quad (3.4.3)$$

where $\mathbf{y} := \text{vec}(Y)$ and $\boldsymbol{\mu}^* := (\mu', \dots, \mu')'$ are $(TK \times 1)$ vectors and $Y_0 := (y'_0, \dots, y'_{-p+1})'$ and $\boldsymbol{\mu} := (\mu', \dots, \mu')'$ are $(Kp \times 1)$. Consequently, $\partial \mathbf{u} / \partial \mathbf{y}'$ is a lower triangular matrix with unit diagonal which has unit determinant. Hence, using that $\mathbf{u} = \mathbf{y} - \boldsymbol{\mu}^* - (X' \otimes I_K)\boldsymbol{\alpha}$,

$$\begin{aligned} f_{\mathbf{y}}(\mathbf{y}) &= \left| \frac{\partial \mathbf{u}}{\partial \mathbf{y}'} \right| f_{\mathbf{u}}(\mathbf{u}) \\ &= \frac{1}{(2\pi)^{KT/2}} |I_T \otimes \Sigma_u|^{-1/2} \\ &\quad \times \exp \left[-\frac{1}{2} (\mathbf{y} - \boldsymbol{\mu}^* - (X' \otimes I_K)\boldsymbol{\alpha})' (I_T \otimes \Sigma_u^{-1}) \right. \\ &\quad \left. \times (\mathbf{y} - \boldsymbol{\mu}^* - (X' \otimes I_K)\boldsymbol{\alpha}) \right], \end{aligned} \quad (3.4.4)$$

where X and $\boldsymbol{\alpha}$ are as defined in (3.3.2). For simplicity, the initial values Y_0 are assumed to be given fixed numbers. Hence, we get a log-likelihood function

$$\ln l(\boldsymbol{\mu}, \boldsymbol{\alpha}, \Sigma_u)$$

Figura 11.11:

$$\begin{aligned}
&= -\frac{KT}{2} \ln 2\pi - \frac{T}{2} \ln |\Sigma_u| \\
&\quad - \frac{1}{2} [\mathbf{y} - \boldsymbol{\mu}^* - (X' \otimes I_K) \boldsymbol{\alpha}]' (I_T \otimes \Sigma_u^{-1}) [\mathbf{y} - \boldsymbol{\mu}^* - (X' \otimes I_K) \boldsymbol{\alpha}] \\
&= -\frac{KT}{2} \ln 2\pi - \frac{T}{2} \ln |\Sigma_u| - \frac{1}{2} \sum_{t=1}^T \left[(y_t - \mu) - \sum_{i=1}^p A_i(y_{t-i} - \mu) \right]' \\
&\quad \times \Sigma_u^{-1} \left[(y_t - \mu) - \sum_{i=1}^p A_i(y_{t-i} - \mu) \right] \\
&= -\frac{KT}{2} \ln 2\pi - \frac{T}{2} \ln |\Sigma_u| \\
&\quad - \frac{1}{2} \sum_t \left(y_t - \sum_i A_i y_{t-i} \right)' \Sigma_u^{-1} \left(y_t - \sum_i A_i y_{t-i} \right) \\
&\quad + \mu' \left(I_K - \sum_i A_i \right)' \Sigma_u^{-1} \sum_t \left(y_t - \sum_i A_i y_{t-i} \right) \\
&\quad - \frac{T}{2} \mu' \left(I_K - \sum_i A_i \right)' \Sigma_u^{-1} \left(I_K - \sum_i A_i \right) \mu \\
&= -\frac{KT}{2} \ln 2\pi - \frac{T}{2} \ln |\Sigma_u| - \frac{1}{2} \text{tr}[(Y^0 - AX)' \Sigma_u^{-1} (Y^0 - AX)], \quad (3.4.5)
\end{aligned}$$

where $Y^0 := (y_1 - \mu, \dots, y_T - \mu)$ and $A := (A_1, \dots, A_p)$ are as defined in (3.3.2). These different expressions of the log-likelihood function will be useful in the following.

3.4.2 The ML Estimators

In order to determine the ML estimators of μ , $\boldsymbol{\alpha}$, and Σ_u , the system of first order partial derivatives is needed:

$$\begin{aligned}
\frac{\partial \ln l}{\partial \mu} &= \left(I_K - \sum_i A_i \right)' \Sigma_u^{-1} \sum_t \left(y_t - \sum_i A_i y_{t-i} \right) \\
&\quad - T \left(I_K - \sum_i A_i \right)' \Sigma_u^{-1} \left(I_K - \sum_i A_i \right) \mu \\
&= [I_K - A(\mathbf{j} \otimes I_K)]' \Sigma_u^{-1} \left[\sum_t (y_t - \mu - AY_{t-1}^0) \right], \quad (3.4.6)
\end{aligned}$$

where Y_t^0 is as defined in (3.3.2) and $\mathbf{j} := (1, \dots, 1)'$ is a $(p \times 1)$ vector of ones,

Figura 11.12:

$$\begin{aligned}\frac{\partial \ln l}{\partial \alpha} &= (X \otimes I_K)(I_T \otimes \Sigma_u^{-1})[\mathbf{y} - \boldsymbol{\mu}^* - (X' \otimes I_K)\boldsymbol{\alpha}] \\ &= (X \otimes \Sigma_u^{-1})(\mathbf{y} - \boldsymbol{\mu}^*) - (XX' \otimes \Sigma_u^{-1})\boldsymbol{\alpha},\end{aligned}\quad (3.4.7)$$

$$\frac{\partial \ln l}{\partial \Sigma_u} = -\frac{T}{2}\Sigma_u^{-1} + \frac{1}{2}\Sigma_u^{-1}(Y^0 - AX)(Y^0 - AX)'\Sigma_u^{-1}. \quad (3.4.8)$$

Equating to zero gives the system of normal equations which can be solved for the estimators:

$$\tilde{\mu} = \frac{1}{T} \left(I_K - \sum_i \tilde{A}_i \right)^{-1} \sum_t \left(y_t - \sum_i \tilde{A}_i y_{t-i} \right), \quad (3.4.9)$$

$$\tilde{\alpha} = ((\tilde{X}\tilde{X}')^{-1}\tilde{X} \otimes I_K)(\mathbf{y} - \tilde{\mu}^*), \quad (3.4.10)$$

$$\tilde{\Sigma}_u = \frac{1}{T}(\tilde{Y}^0 - \tilde{A}\tilde{X})(\tilde{Y}^0 - \tilde{A}\tilde{X})', \quad (3.4.11)$$

where \tilde{X} and \tilde{Y}^0 are obtained from X and Y^0 , respectively, by replacing μ with $\tilde{\mu}$.

Figura 11.13:

con la definición que tenemos en seguida

$$\begin{aligned}Y^0 &:= (y_1 - \mu, \dots, y_T - \mu) \quad (K \times T), \\ A &:= (A_1, \dots, A_p) \quad (K \times Kp), \\ Y_t^0 &:= \begin{bmatrix} y_t - \mu \\ \vdots \\ y_{t-p+1} - \mu \end{bmatrix} \quad (Kp \times 1), \\ X &:= (Y_{-1}^0, \dots, Y_{T-1}^0) \quad (Kp \times T), \\ \mathbf{y}^0 &:= \text{vec}(Y^0) \quad (KT \times 1), \\ \alpha &:= \text{vec}(A) \quad (K^p p \times 1),\end{aligned}\quad (3.3.2)$$

we can write (3.3.1), for $t = 1, \dots, T$, compactly as

$$Y^0 = AX + U \quad (3.3.3)$$

or

$$\mathbf{y}^0 = (X' \otimes I_K)\boldsymbol{\alpha} + \mathbf{u}, \quad (3.3.4)$$

Figura 11.14:

Propiedades de los Estimadores de Máxima Verosimilitud.

Al comparar las formas que tienen los estimadores de máxima verosimilitud y los de mínimos cuadrados, podemos verificar que los estimadores son idénticos. Adicionalmente tenemos la siguiente proposición:

Proposición 11.42. *Propiedades Asintóticas de los Estimadores de Máxima Verosimilitud*

Let y_t be a stationary, stable Gaussian VAR(p) process as in (3.3.1). Then the ML estimators $\tilde{\mu}$, $\tilde{\alpha}$, and $\tilde{\sigma} = \text{vech}(\tilde{\Sigma}_u)$ given in (3.4.9)–(3.4.11) are consistent and

$$\sqrt{T} \begin{bmatrix} \tilde{\mu} - \mu \\ \tilde{\alpha} - \alpha \\ \tilde{\sigma} - \sigma \end{bmatrix} \xrightarrow{d} \mathcal{N} \left(0, \begin{bmatrix} \Sigma_{\tilde{\mu}} & 0 & 0 \\ 0 & \Sigma_{\tilde{\alpha}} & 0 \\ 0 & 0 & \Sigma_{\tilde{\sigma}} \end{bmatrix} \right), \quad (3.4.29)$$

so that $\tilde{\mu}$ is asymptotically independent of $\tilde{\alpha}$ and $\tilde{\Sigma}_u$ and $\tilde{\alpha}$ is asymptotically independent of $\tilde{\mu}$ and $\tilde{\Sigma}_u$. The covariance matrices are

$$\begin{aligned} \Sigma_{\tilde{\mu}} &= \left(I_K - \sum_i A_i \right)^{-1} \Sigma_u \left(I_K - \sum_i A'_i \right)^{-1}, \\ \Sigma_{\tilde{\alpha}} &= \Gamma_Y(0)^{-1} \otimes \Sigma_u, \\ \Sigma_{\tilde{\sigma}} &= 2\mathbf{D}_K^+(\Sigma_u \otimes \Sigma_u)\mathbf{D}_K^{+\prime}. \end{aligned}$$

They may be estimated consistently by replacing the unknown quantities by their ML estimators and estimating $\Gamma_Y(0)$ by $\tilde{X}\tilde{X}'/T$. ■

Figura 11.15:

11.4 Pronóstico con Procesos Estimados

Las expresiones encontradas para el predictor óptimo se dieron asumiendo que los verdaderos valores de los parámetros son conocidos. Sin embargo, en la práctica uno tiene es una estimación de estos \hat{B} , y por lo tanto el pronóstico puede verse afectado. En el caso del modelo VAR(p), si usamos \hat{B} , tenemos que el predictor óptimo queda

$$\hat{Y}_t(h) = \hat{\nu} + \hat{A}_1 \hat{Y}_t(h-1) + \cdots + \hat{A}_p \hat{Y}_t(h-p) \quad (11.24)$$

donde $\hat{Y}_t(j) = \hat{Y}_{t+j}$ para $j \leq 0$.

Con los cuales el error de pronóstico tiene la siguiente expresión:

$$\begin{aligned} \hat{Y}_{t+h} - \hat{Y}_t(h) &= [\hat{Y}_{t+h} - \hat{Y}_t(h)] + [\hat{Y}_t(h) - \hat{Y}_t(h)] \\ &= \sum_{i=0}^{h-1} \Phi_i \hat{u}_{t+h-i} + [\hat{Y}_t(h) - \hat{Y}_t(h)]. \end{aligned}$$

Se puede verificar que este estimador óptimo también es insesgado. Ahora, el error cuadrático medio de pronóstico con los parámetros estimados tiene la siguiente expresión:

$$\Sigma_{\hat{Y}}(h) = MSE[\hat{Y}_t(h)] = \Sigma_{\hat{Y}}(h) + MSE[\hat{Y}_t(h) - \hat{Y}_t(h)]$$

con $\Sigma_{\hat{Y}} = \sum_{i=0}^{h-1} \Phi_i \sigma_u \Phi'_i$.

con esto, se requiere ver como estimar $MSE[\hat{Y}_t(h) - \hat{\tilde{Y}}_t(h)]$. Se puede verificar que una aproximación está dada por $\Omega(h)/T$, donde

$$\Omega(h) = E \left[\frac{\partial \hat{Y}_t(h)}{\partial \beta'} \Sigma_{\hat{\beta}} \frac{\partial \hat{Y}_t(h)}{\partial \beta} \right].$$

Así que una aproximación tenemos

$$\Sigma_{\hat{Y}}(h) = \Sigma_{\hat{Y}}(h) + \frac{1}{T} \Omega(h).$$

Nota 11.43. Note que por la propiedad que tienen los estimadores de MC o MV, de ser consistentes, uno puede asumir que $\hat{B} \approx B$, cuando $T \rightarrow \infty$, y así los resultados (con parámetros verdaderos y los estimadores) son asintóticamente equivalentes.

Cómo se obtiene una encontrar una expresión explícita para $\Omega(h)$? Leer sección 3.5.2 del libro [27].

11.5 Selección y Diagnóstico del Modelo

11.5.1 Criterios de Selección del Modelo

Parte de esta sección se llevada a cabo teniendo en cuenta el libro [37]. La idea consiste en usar pruebas de cociente de verosimilitud secuenciales , criterios de información o minimizar el error cuadrático medio de predicción. Veamos primero las pruebas de cociente de verosimilitud.

Pruebas secuenciales de cociente de verosimilitud

La idea básica de este enfoque es comparar un modelo $VAR(\ell)$ con el modelo $VAR(\ell - 1)$. Desde un enfoque estadístico lo que hay que probar es

$$H_0 : A_\ell = \mathbf{0} \quad v.s \quad H_a : A_\ell \neq \mathbf{0}.$$

Este es un problema de hipótesis anidadas, el cual puede resolverse usando el estadístico de cociente de verosimilitud. En general, el estadístico de cociente de verosimilitud consiste

$$\lambda_{LR} = 2[\ln l(\tilde{\delta}) - \ln l(\tilde{\delta}_r)]$$

donde ($\tilde{\delta}$ es el estimador de máxima verosimilitud sin restringir para el parámetro $\tilde{\delta}$, y $\tilde{\delta}_r$ es el estimador de ML restringido, es decir, el que se obtiene maximizando la función de verosimilitud sobre esa parte del espacio de parámetros donde las restricciones de interés se satisfacen. Para el caso de modelos VAR, donde se tienen restricciones lineales para los coeficientes del modelo, la distribución asintótica de λ_{LR} es una χ^2 con tantos grados de libertad, como restricciones lineales distintas se tengan.

Para llevar a cabo este procedimiento, consideremos que la matriz de coeficientes $\beta'_\ell = [\underline{y}, A_1, \dots, A_\ell]$ y $\Sigma_{\underline{y}, \ell}$, la matriz de covarianza de los errores o innovaciones. El estadístico de cociente de verosimilitud que es usado es

$$M(\ell) = -(T - \ell - 1.5 + K\ell) \ln \left(\frac{|\hat{\Sigma}_{\underline{y}, \ell}|}{|\hat{\Sigma}_{\underline{y}, \ell-1}|} \right)$$

el cual tiene distribución χ^2 con K^2 grados de libertad. Taio y Box en 1980 sugieren el siguiente procedimiento para computar el estadístico $M(\ell)$ y seleccionar el orden del modelo VAR:

1. Seleccione un entero positivo P , el cual es el orden máximo del modelo que será entrenado.
2. Configure la regresión lineal multivariada basada en el modelo 11.19, transponiendo todo para el modelo inicial $VAR(P)$,
3. Para $\ell = 0, \dots, P$, compute la estimación de mínimos cuadrados para la matriz de coeficientes AR, es decir, compute $\hat{\beta}_\ell$. Luego, compute la estimación de MV de $\Sigma_{\underline{y}}$, es decir, compute $\hat{\Sigma}_{\underline{y}, \ell} = \frac{1}{T-P}(Y - \hat{B}_\ell Z)(Y - \hat{B}_\ell Z)'$.
4. Para $\ell = 0, \dots, P$, compute el estadístico de prueba cociente de verosimilitud:

$$M(\ell) = -(T - P - 1.5 + K\ell) \ln \left(\frac{|\hat{\Sigma}_{\underline{y}, \ell}|}{|\hat{\Sigma}_{\underline{y}, \ell-1}|} \right)$$

y su respectivo valor $-p$, el cual está basado en la distribución asintótica $\chi^2_{K^2}$.

5. Examine los estadísticos de prueba secuencialmente empezando con $\ell = 1$. Si todos los valores p de las estadísticas $M(\ell)$ son mas grandes que el valor de error tipo I para $\ell > p$, entonces un modelo $VAR(p)$ es especificado. Esto es así porque la prueba rechaza las hipótesis nulas para A_p , pero falla en rechazar para $A_\ell = \mathbf{0}$ para $\ell > p$.

Criterios de Información

Los tres criterios que son mas comúnmente usados para determinar el orden de un modelo $VAR(\ell)$ son los siguientes bajo el supuesto de normalidad:

$$\begin{aligned} AIC(\ell) &= \ln |\hat{\Sigma}_{\underline{u},\ell}| + \frac{2}{T} \ell K^2, \\ BIC(\ell) &= \ln |\hat{\Sigma}_{\underline{u},\ell}| + \frac{\ln(T)}{T} \ell K^2, \\ HQ(\ell) &= \ln |\hat{\Sigma}_{\underline{u},\ell}| + \frac{2 \ln[\ln(T)]}{T} \ell K^2. \end{aligned}$$

El criterio de BIC y HQ son consistentes en el sentido que selecciona el verdadero $VAR(p)$ c.p.1 cuando $T \rightarrow \infty$. Otro criterio, el cual se conoce como el error de predicción final(FPE), es el está basado en una estimación del error de predicción un paso adelante, el cual se define como

$$\begin{aligned} FPE(\ell) &= \det \left[\frac{T + K\ell + 1}{T} \frac{T}{T - K\ell - 1} \tilde{\Sigma}_{\underline{u},\ell} \right] \\ &= \left[\frac{T + K\ell + 1}{T - K\ell - 1} \right]^K \det \tilde{\Sigma}_{\underline{u},\ell}, \end{aligned}$$

lo cual permite escoger el orden p tal que se minimice el FPE .

Análisis de Residuales

Vale la pena señalar, que el diagnóstico de residuales es un etapa importante en la construcción del modelo. Esto le permite asegurar que el modelo ajustado es adecuado y sugerir como se podría mejorar el modelo. Se dice que un modelo es adecuado si a) todos los parámetros son significativos, b) los residuales no presentan correlación serial o autocorrelación correlación cruzada, c) no existen cambios estructurales o outliers, d) los residuales no violan el supuesto distribucional. Veamos si las correlaciones cruzadas de los errores basados en los residuales del modelo:

$$\hat{u}_t = \underline{Y}_t - \hat{\nu} + \sum_{i=1}^p \hat{A}_i \underline{Y}_{t-i}.$$

Correlaciones Cruzadas de los Residuales

Sea R_j la matriz de correlación cruzada de rezagos j del proceso $\{\underline{u}_t\}$ de un modelo VAR(p). Con el objeto de hacer le chequeo del modelo, la hipótesis nula de interés será

$$H_0 : R_1 = R_2 = \cdots = R_m = \mathbf{0} \quad v.s. \quad H_a : R_i \neq \mathbf{0} \quad \text{para algún } 1 \leq i \leq m$$

para un entero m pre-especificado. Sea \hat{u}_t la serie de residuales. Un estimador natural de las matrices R_j es

$$\hat{R}_j = \hat{D}^{-1/2} \hat{C}_j \hat{D}^{-1/2}$$

con \hat{C}_j la matriz de autocovarianza matricial de rezago j dada por

$$\hat{C}_j = \frac{1}{T-p} \sum_{t=p+j+1}^T \hat{u}_t \hat{u}'_{t-j}, \quad \text{para } j = 0, 1, \dots,$$

y $\hat{D} = \text{diag}\{\hat{C}_{0,11}, \dots, \hat{C}_{0,KK}\}$ de los elementos de la diagonal de la matrix \hat{C}_0 . Con esto, se puede verificar que $\hat{\varepsilon}_m = [\hat{R}_1, \dots, \hat{R}_m]$

$$\sqrt{T_p} \text{vec}(\hat{\varepsilon}_m) \xrightarrow[T_p \rightarrow \infty]{d} N(\mathbf{0}, \Sigma_{r,m}).$$

Con esto, se puede verificar que el estadístico portmanteau multivariado definido como

$$\begin{aligned} Q_k(m) &= T^2 \sum_{\ell=1}^m \frac{1}{T-\ell} \text{tr} \left(\hat{R}'_\ell \hat{R}_0^{-1} \hat{R}_\ell \hat{R}_0^{-1} \right) \\ &= T^2 \sum_{\ell=1}^m \frac{1}{T-\ell} \text{tr} \left(\hat{C}'_\ell \hat{C}_0^{-1} \hat{C}_\ell \hat{C}_0^{-1} \right) \end{aligned}$$

tiene distribución χ_d^2 con $d = (m-p)K^2$. Con estos resultados, también se puede chequear si las autocorrelaciones individuales son significativas, comparando sus estimaciones como sigue $|r_{mn,i}| > 2/\sqrt{T}$.

Nota 11.44. Recuerde que cerca de 1 de cada 20 autocorrelaciones serán significativas aún cuando el verdadero proceso subyacente sea un ruido blanco.

Prueba de Parámetros Cero

En ocasiones se requiere remover parámetros que no son significativos. Para hacer esto al menos dos enfoques, por favor leer del libro [37] las páginas 73 a 76. Usar las funciones *refVAR* y *VARchi*.

Prueba de No-Normalidad

Vamos a ilustrar como llevar a cabo la prueba de Jarque-Bera de normalidad multivariada. Para esto requerimos la siguiente proposición:

Proposición 11.45. *Distribución Asintótica de la Curtosis y la Asimetría Residual*

Let \hat{y}_t be a K -dimensional stationary, stable Gaussian VAR(p) process as in (4.5.10), where u_t is zero mean white noise with nonsingular covariance matrix Σ_u and let $\hat{A}_1, \dots, \hat{A}_p$ be consistent and asymptotically normally distributed estimators of the coefficients based on a sample y_1, \dots, y_T and possibly some presample values. Define

$$\hat{u}_t := (y_t - \bar{y}) - \hat{A}_1(y_{t-1} - \bar{y}) - \dots - \hat{A}_p(y_{t-p} - \bar{y}), \quad t = 1, \dots, T,$$

$$\hat{\Sigma}_u := \frac{1}{T - Kp - 1} \sum_{t=1}^T \hat{u}_t \hat{u}'_t,$$

and let \hat{P} be a matrix satisfying $\hat{P}\hat{P}' = \hat{\Sigma}_u$ such that $\text{plim}(\hat{P} - P) = 0$. Furthermore, define

$$\hat{w}_t = (\hat{w}_{1t}, \dots, \hat{w}_{Kt})' := \hat{P}^{-1}\hat{u}_t,$$

$$\hat{b}_1 = (\hat{b}_{11}, \dots, \hat{b}_{K1})' \quad \text{with} \quad \hat{b}_{k1} := \frac{1}{T} \sum_{t=1}^T \hat{w}_{kt}^3, \quad k = 1, \dots, K,$$

and

$$\hat{b}_2 = (\hat{b}_{12}, \dots, \hat{b}_{K2})' \quad \text{with} \quad \hat{b}_{k2} := \frac{1}{T} \sum_{t=1}^T \hat{w}_{kt}^4, \quad k = 1, \dots, K.$$

Then

$$\sqrt{T} \begin{bmatrix} \hat{b}_1 \\ \hat{b}_2 - \mathbf{3}_K \end{bmatrix} \xrightarrow{d} \mathcal{N} \left(0, \begin{bmatrix} 6I_K & 0 \\ 0 & 24I_K \end{bmatrix} \right).$$

■

Figura 11.16:

Basado en la proposición anterior, se puede verificar que los estadísticos

$$\hat{\lambda}_s = T\hat{b}'_1\hat{b}_1/6 \rightarrow \chi^2(K)$$

$$\hat{\lambda}_k = T(\hat{b}_2 - \mathbf{3}_K)'(\hat{b}_2 - \mathbf{3}_K)/24 \rightarrow \chi^2(K),$$

$$\hat{\lambda}_{sk} = \hat{\lambda}_s + \hat{\lambda}_k \rightarrow \chi^2(2K).$$

Esto permite llevar a cabo la prueba de normal multivariada basada en los residuales del modelo.

Pruebas para Cambios Estructurales

El supuesto de estabilidad o estacionariedad es importante en el modelamiento. Sin embargo, en ocasiones eventos pueden causar turbulencia en los sistemas económicos, por ejemplo las guerras pueden cambiar los valores de la variable de interés. También, las políticas acerca de los impuestos o tasas de interés pueden generar un impacto en la economía. Se pueden usar pruebas del tipo *Chow* para chequear si hay cambio en los parámetros ha ocurrido en un tiempo t , comparando los parámetros estimados antes y después del tiempo t . Para esto se pueden usar pruebas del tipo *Wald*

o del tiempo cociente de verosimilitud. También se pueden usar de forma univariadas las estadísticas Cusum y Cusumsq.

11.6 Prueba de Causalidad de Granger

11.6.1 Prueba de Causalidad de Granger del tipo Wald

Anteriormente, se dedujo que para dado un proceso $\underline{Y}'_t = (\underline{Z}'_t, \underline{X}'_t)$ VAR(p), la no causalidad de Granger estaba caracterizada por 11.14, mas específicamente, por las siguientes restricciones:

$$\begin{aligned} \underline{Z}_t(1|\{\underline{Y}_s|s \leq t\}) &= \underline{Z}_t(1|\{\underline{Z}_s|s \leq t\}), \quad h = 1, 2, \dots, \\ \Leftrightarrow A_{12,i} &= 0 \text{ para } i = 1, 2, \dots, p. \end{aligned}$$

o alternativamente,

$$\begin{aligned} \underline{X}_t(h|\{\underline{Y}_s|s \leq t\}) &= \underline{X}_t(h|\{\underline{X}_s|s \leq t\}), \quad h = 1, 2, \dots, \\ \Leftrightarrow A_{21,i} &= 0 \text{ para } i = 1, 2, \dots, p. \end{aligned}$$

Para probar estas restricciones, se requiere de los resultados de las propiedades de los estimadores dadas anteriormente. En forma general, se considera probar

$$H_0 : C\beta = \underline{c} \text{ v.s. } H_1 : C\beta \neq \underline{c} \quad (11.25)$$

donde C es una matriz $X \times (K^2p + K)$ de rango N y \underline{c} es un vector $N \times 1$. Asumiendo que

$$\sqrt{T}(\hat{\beta} - \beta) \xrightarrow{d} N(\underline{0}, \Gamma^{-1} \otimes \Sigma_u)$$

as in LS/ML estimation, we get

$$\sqrt{T}(C\hat{\beta} - C\beta) \xrightarrow{d} N[0, C(\Gamma^{-1} \otimes \Sigma_u)C'] \quad (3.6.3)$$

(see Appendix C, Proposition C.15) and, hence,

$$T(C\hat{\beta} - c)' [C(\Gamma^{-1} \otimes \Sigma_u)C']^{-1} (C\hat{\beta} - c) \xrightarrow{d} \chi^2(N). \quad (3.6.4)$$

This statistic is the *Wald statistic* (see Appendix C.7).

Figura 11.17:

Con los cual reemplazando a Σ y Σ_u por sus respectivos estimadores, tenemos que el estadístico

$$\lambda_W = \left(C\hat{\beta} - c \right)' \left[C((ZZ')^{-1} \mathbf{1} \otimes \hat{\Sigma}_u)C' \right]^{-1} \left(C\hat{\beta} - c \right)$$

tiene distribución asintótica χ^2_N . Con los cual tenemos la siguiente proposición

Proposición 11.46. (*Distribución asintótica del estadístico de Wald*)

Suppose (3.6.2) holds. Furthermore, $\text{plim}(ZZ'/T) = \Gamma$, $\text{plim } \hat{\Sigma}_u = \Sigma_u$ are both nonsingular and $H_0 : C\beta = c$ is true, with C being an $(N \times (K^2p + K))$ matrix of rank N . Then

$$\lambda_W = (C\hat{\beta} - c)' [C((ZZ')^{-1} \otimes \hat{\Sigma}_u)C']^{-1} (C\hat{\beta} - c) \xrightarrow{d} \chi^2(N).$$

■

Figura 11.18:

Usualmente se suele trabajar con una variante este estadístico

$$\hat{\lambda}_F = \frac{\hat{\lambda}_W}{N}$$

donde N es el rango de la matriz C . La cual tiene una distribución aproximadamente como una $F(N, T - Kp - 1)$.

12

Procesos Integrados y Modelos de Corrección de Errores Vectoriales

Vale la pena recordar que una series de tiempo escalar $\{Z_t\}$ tiene una raíz unitaria de orden 1 si $W_t = (1 - B)Z_t$ es proceso estacionario e invertible, así, el proceso ejse conoce como $I(1)$, o integrado de orden 1. En forma general un proceso $\{Z_t\}$ es un proceso $I(d)$ si $W_t = (1 - B)^d Z_t$ es estacionario e invertible.

Definición 12.1. *Un proceso multivariado k -dimensional Z_t definido como*

$$\Phi(B)Z_t = \zeta_0 + \Theta(B)\alpha_t$$

con $\Phi(B) = I - \sum_{i=1}^p \Phi_i(B)$ y $\Theta(B) = I - \sum_{i=1}^q \Theta_i(B)$, con p, q los órdenes de los polinomios matriciales. Diremos que el proceso es integrado si $|\Phi(1)| = 0$, pero $|\Theta(1)| \neq 0$, es decir, "1" es una solución de la ecuación determinante $|\Phi(z)| = 0$, pero no de $|\Theta(z)| = 0$.

Los procesos integrados son usados en muchas aplicaciones y en especial los $I(1)$. Sin embargo, hay que tener cuidado con las relaciones espurias.

12.1 Regresión Espuria

Vamos a suponer que los procesos Z_{1t} y Z_{2t} son procesos $I(1)$ escalares, y se desea saber si existe alguna relación lineal entre los procesos. Para esto se propone la regresión:

$$Z_{2t} = \alpha + \beta Z_{1t} + \epsilon_t$$

donde $\{\epsilon_t\}$ denota el proceso de error. Si $\beta \neq 0$ entonces los dos procesos están linealmente correlacionados, y entonces para probar la hipótesis nula $H_0 : \beta = 0$ v.s. $H_a : \beta \neq 0$, no debería usarse la tradicional distribución del estadístico cociente t basado en el estimador de MC para β . Si se llegara usar este estadístico, uno estaría haciendo descubrimientos que son falsos, es decir, rechazando muy frecuentemente H_0 cuando esta es verdadera. Usted puede realizar el siguiente ejercicio de simulación para que chequee lo anterior:

- ✓ Simule dos procesos caminatas aleatorias de forma independiente

$$Z_{1t} = Z_{1,t-1} + a_{1t}, \quad z_{10} = 0$$

$$Z_{2t} = Z_{2,t-1} + a_{2t}, \quad z_{20} = 0, \quad t = 1, \dots, 1000.$$

siendo $\{a_{1t}\}$ y $\{a_{2t}\}$ procesos Gaussianos independientes.

- ✓ Ajuste la regresión propuesta anteriormente entre Z_{2t} y Z_{1t} .
- ✓ Use el valor crítico usual de 1.96, y recolectemos cuantas veces se rechaza la hipótesis nula.
- ✓ Repita el procedimiento anterior 1000 réplicas

Se espera que aproximadamente el 77% de las veces se rechace la hipótesis nula, lo cual es mas alto que el 5% del error tipo I. Como moraleja, uno debe tener cuidado cuando se buscan relaciones lineales en procesos integrados. Es decir, la prueba puede encontrar relaciones que no existen cuando se manejan procesos $I(1)$, es decir hay relaciones espurias. Para lo cual se debe chequear si los residuales de la regresión son $I(1)$, si esto es así, entonces la relación es espuria.

12.2 Combinaciones Lineales de un Proceso Vectorial

Las combinaciones lineales de un vector de series de tiempo son de la forma: $Y_t = \underline{b}' \underline{Z}_t$, y puede conducir a algunos procesos escalares interesantes con aplicaciones importantes. Es importante notar que si se elige a \underline{b} como los valores propios de la matriz de covarianza de \underline{Z}_t , se obtiene una componente principal. Vale la pena decir, que el PCA descompone a \underline{Z}_t dentro de un conjunto de series no correlacionadas ordenadas por las magnitudes de sus varianzas. Existe otro enfoque que consiste en encontrar las combinaciones lineales de \underline{Z}_t basados en sus predictibilidades un paso adelante el

cual da como resultado que la combinación lineal con máxima predictibilidad es equivalente a la variable canónica correspondiente a la correlación canónica mas grande entre \underline{Z}_t y su pasado(es decir, el vector de que genera la combinación lineal es el vector propio de la matriz de covarianza canónica $M = \Gamma_z^{-1}(0)\Gamma_z(1)\Gamma_z^{-1}(0)\Gamma_z(1)'$.

12.3 Co-integración

Un fenómeno importante de los procesos integrados es la posibilidad de co-integración. Considere los procesos univariados $I(1)$, $\{Z_{1t}\}$ y Z_{2t} . Si existe una combinación lineal diferente a la nula $Y_t = \beta_1 Z_{1t} + \beta_2 Z_{2t}$, la cual es estacionaria, entonces diremos que Z_{1t} y Z_{2t} están co-integradas. el vector $\beta = (\beta_1, \beta_2)'$ hace referencia al vector de co-integración. En forma general tenemos la siguiente definición:

Definición 12.2. *Sea $\{Z_t\}$ un proceso k -dimensional integrado. Si existe un vector no-nulo β k -dimensional, tal que $Y_t = \beta' \underline{Z}_t$ es estacionario o $I(0)$, entonces diremos que $\{Z_t\}$ es llamado que es co-integrado con vector de cointegración β .*

Nota 12.3. *Si existe una matriz $\beta, K \times m$ de rango completo m tal que el proceso m -dimensional $\underline{Y}_t = \beta' \underline{Z}_t$ es estacionario, entonces diremos que $\{Z_t\}$ es co-integrado con rango de co-integración m y matriz de co-integración β . Es decir, m es el número de vectores de co-integración o relaciones de co-integración linealmente independientes.*

Nota 12.4. *Vale la pena decir que la idea de co-integración en economía y finanzas está relacionada al concepto del equilibrio de largo plazo en economía y a la idea de comercio de pares en finanzas. Con esto en mente, la co-integración quiere decir que mientras que las componentes individuales de \underline{Z}_t no son predecibles, sin embargo sus combinaciones lineales $\underline{Y}_t = \beta' \underline{Z}_t$ si lo son.*

12.3.1 Sobre-diferenciación

Usualmente cuando se manipulan series de tiempo no estacionarias con raíz unitaria, se procede primero a tomar diferencias. Por definición, si una serie escalar es un proceso $I(1)$, entonces $W_t = (1 - B)Z_t$ es estacionario e invertible. Por lo tanto, se puede analizar el proceso W_t y producir el pronóstico un paso adelante $\hat{Z}_T(1) = E(Z_{T+1}|F_T) = Z_T + \hat{w}_T(1)$, sin embargo para el caso

multivariado las cosas son mas complejas. Si $\{Z_t\}$ es un proceso co-integrado de rango de co-integración m con $m < k$, entonces existen únicamente $k - m$ raíces unitarias en el proceso k -dimensional Z_t . Si uno considera entonces $W_t = (1 - B)Z_t$, es decir, $W_{it} = Z_{it} - Z_{it-1}$, W_t puede no ser invertible. Veamos un ejemplo de esto. Considere el proceso bivariado $ARMA(1, 1)$

$$Z_t - \begin{bmatrix} 1.05 & -0.05 \\ 0.45 & 0.55 \end{bmatrix} Z_{t-1} = \underline{a}_t - \begin{bmatrix} -0.05 & 0.45 \\ 0.45 & -0.05 \end{bmatrix} \underline{a}_{t-1}$$

donde \underline{a}_t es un proceso de ruido blanco bi-dimensional con matriz de covarianza Σ_a . Note que el determinante de los operadores son

$$|\Phi(z)| = (1 - z)(1 - 0.6z), \quad |\Theta(z)| = (1 - 0.4z)(1 + 0.5z)$$

lo cual permite ver que tiene una raíz unitaria, pero es invertible, es decir $|\Phi(1)| = 0$ pero $|\Theta(1)| \neq 0$. Ademas, se puede verificar que

$$Z_{1t} - Z_{2t} = 0.6(Z_{1,t-1} - Z_{2,t-1}) + (a_{1t} - a_{2t}) + 0.5(a_{1,t-1} - a_{2,t-1}).$$

ahora, estableciendo que $X_t = Z_{1t} - Z_{2t}$ y $b_t = a_{1t} - a_{2t}$, tenemos que

$$X_t = 0.6X_{t-1} + b_t + 0.5b_{t-1}$$

es estacionario e invertible, por qué? Por lo tanto, $Z_{1,t}$ y $Z_{2,t}$ son co-integrados y el vector de co-integración es $(1, -1)'$.

Luego, sea $\underline{W}_t = (1 - B)Z_t$, es decir, tomar las primeras diferencias de cada serie. Ahora, pre-multiplicando el proceso $VARMA(1, 1)$ por

$$\begin{bmatrix} 1 - 0.55B & -0.055B \\ 0.45B & 1 - 1.05B \end{bmatrix}$$

se obtiene(haciendo operaciones)

$$(1 - 0.6B)(1 - B)I_2Z_t = \begin{bmatrix} 1 - 0.5B - 0.005B^2 & -0.5B + 0.245B^2 \\ 0.495B^2 & 1 - B - 0.255B^2 \end{bmatrix} \underline{a}_t$$

Por ejemplo, note que el lado izquierdo se obtiene de la siguiente forma:

$$\begin{bmatrix} 1 - 0.55B & -0.055B \\ 0.45B & 1 - 1.05B \end{bmatrix} Z_t - \begin{bmatrix} 1 - 0.55B & -0.055B \\ 0.45B & 1 - 1.05B \end{bmatrix} \begin{bmatrix} 1.05B & -0.05B \\ 0.45B & 0.55B \end{bmatrix} Z_t,$$

así, podemos ver que la primera componente de la matriz resultante es obtenida como sigue:

$$1 - 0.55B - [(1 - 0.55B)1.05B - (0.05B)(0.45B)] = 1 - 0.55B - 10.5B + 0.6B^2 = 1 - 1.6B + 0.6B^2 = (1 - B)(1 - 0.6B),$$

mientras que las demás componentes se obtienen de la misma forma. Con lo cual tenemos

$$(1 - 0.6B)\underline{W}_t = \begin{bmatrix} 1 - 0.5B - 0.005B^2 & -0.5B + 0.245B^2 \\ 0.495B^2 & 1 - B - 0.255B^2 \end{bmatrix} \underline{a}_t,$$

es decir el proceso $\{\underline{W}_t\}$ es un proceso VARMA(1,2). Sin embargo, se puede verificar que

$$|\Theta^*(1)| = \begin{vmatrix} 1 - 0.5 - 0.005 & -0.5 + 0.245 \\ 0.495 & 1 - 1 - 0.255 \end{vmatrix} = \begin{vmatrix} 0.495 & -0.255 \\ 0.495 & -0.255 \end{vmatrix} = 0.$$

Es decir, el proceso $\{\underline{W}_t\}$ no es invertible. Con lo cual, si se diferencia componentes individuales de un proceso co-integrado vectorial con rango de co-integración mas pequeño que la dimensión del vector de respuesta, resulta en un modelo VARMA no-invertible se conoce como sobre-diferenciación en la literatura de series de tiempo. Esto tiene implicaciones en el modelamiento de procesos coi-integrados, por ejemplo, al no ser invertible, pues no tiene aproximación *VAR* finita, implicando que se requiere un orden alto del modelo *VAR* para alcanzar una buena aproximación. Además, las propiedades de los estimadores de máxima verosimilitud pueden no aplicar en proceso no invertibles. Una solución a esto, es considerará el modelo de corrección de errores.

12.4 Modelos de Corrección de Errores Vectoriales(VEC)

Considere el proceso VARMA

$$\Phi(B)\underline{Z}_t = \underline{\zeta}_0 + \Theta(B)\underline{a}_t \quad (12.1)$$

asumiendo que 1 puede ser una solución de la ecuación determinante $|\Phi(z)| = 0$, es decir, asuma que el proceso $\{\underline{Z}_t\}$ es un proceso que puede ser integrado de orden 1. Sea $\nabla \underline{Z}_t = Z_t - Z_{t-1} = (1 - B)Z_t$, la primera diferencia de Z_t . Se puede verificar que(ver [37] Página 283)

$$\nabla \underline{Z}_t = \Pi \underline{Z}_t + \sum_{i=1}^{p-1} \Phi_i^* \nabla \underline{Z}_{t-i} + \underline{\zeta}_0 + \Theta(B)\underline{a}_t \quad (12.2)$$

donde $\Pi = \sum_{i=1}^p \Phi_i - I = -\Phi(1)$ y $\Phi_i^* = -(\Phi_{i+1} + \dots + \Phi_p)$ para $i = 1, \dots, p-1$.

El modelo 12.2 se conoce como modelo de corrección de errores (VEC o ECM) para el proceso integrado \tilde{Z}_t . Es claro que este modelo es invertible, ya que no se ha alterado el polinomio de promedios móviles matricial no se altera. Por supuesto, si $|\Phi(1)| = 0$, entonces la matriz de coeficientes Π es singular. En general se tiene los siguientes casos:

- ✓ Si $rango(\Pi) = 0$, entonces $\Phi = \mathbf{0}$ y $\{\tilde{Z}_t\}$ tiene k raíces unitarias. Por lo tanto el proceso no es co-integrado.
- ✓ Si $rango(\Pi) = k$, entonces Φ es no singular y $\{\tilde{Z}_t\}$ es un proceso estacionario.
- ✓ Si $rango(\Pi) = m$ con $1 \leq m < k$, entonces $\{\tilde{Z}_t\}$ tiene $v = k - m$ raíces unitarias y existe m vectores de cointegración linealmente independientes. En este caso, se puede escribir $\Pi = \boldsymbol{\alpha}\boldsymbol{\beta}'$, con $\boldsymbol{\alpha}$ y $\boldsymbol{\beta}$ son matrices $k \times m$ reales de rango m . Sea $\underline{W}_t = \boldsymbol{\beta}'\tilde{Z}_t$ un proceso m -dimensional el cual es estacionario. En este caso, las columnas de la matriz $\boldsymbol{\beta}$ son los vectores co-integración de \tilde{Z}_t . También, note que $\boldsymbol{\alpha}\boldsymbol{\beta}' = (\boldsymbol{\alpha}\mathbf{P})(\mathbf{P}\boldsymbol{\beta})'$ para cualquier matriz ortogonal \mathbf{P} $m \times m$. Es decir, los vectores de co-integración no son definidos con unicidad cuando $m > 1$, pero el espacio de columnas de $\boldsymbol{\beta}$ si es único.

Nota 12.5. Note que la estacionariedad del proceso \underline{W}_t de un proceso co-integrado $\{\tilde{Z}_t\}$ es fácilmente deducida re-escribiendo el ECM como sigue

$$\nabla \tilde{Z}_t = \boldsymbol{\alpha} \underline{W}_{t-1} + \sum_{i=1}^{p-1} \Phi_i^* \nabla \tilde{Z}_{t-i} + \underline{c}_0 + \Theta(B) \underline{a}_t$$

con $\boldsymbol{\alpha}$ de rango completo. Es claro que al ser $\nabla \tilde{Z}_t$ estacionario, no puede depender de un proceso con raíz unitaria, así que $\{\underline{W}_t\}$ debe ser estacionario.

Una forma alternativa de escribir el modelo 12.2 es como sigue:

$$\nabla \tilde{Z}_t = \Pi \tilde{Z}_{t-p} + \sum_{i=1}^{p-1} \Phi_i^* \nabla \tilde{Z}_{t-i} + \underline{c}_0 + \Theta(B) \underline{a}_t \quad (12.3)$$

donde Π se mantiene, es decir $\Pi = -\Phi(1)$, pero $\Phi_i^* = -(I - \Phi_1 - \dots - \Phi_i)$ para $i = 1, \dots, p-1$. Las matrices Φ_i^* contienen los efectos acumulados de largo plazo, y el ECM se refiere al ECM de la forma de largo plazo.

12.5 Prueba de Co-integración

Detectar la existencia y así el número del vectores de co-integración de un proceso multivariado se conoce como la *prueba de co-integración* en la literatura de econometría. No hay único enfoque para llevar a cabo estas prueba. Si tenemos un VAR puro, la prueba de co-integración de *Johansen* es mas usada. Para llevar a cabo esta prueba, se usa la escritura 12.2 y 12.3 y se quita el polinomio de promedios móviles matricial, inclusive se puede incluir un término de tendencia \underline{d}_t que depende del tiempo. Esto es,

$$\nabla \underline{Z}_t = \Pi \underline{Z}_{t-p} + \sum_{i=1}^{p-1} \Phi_i^* \nabla \underline{Z}_{t-i} + \underline{d}_t + \underline{a}_t \quad (12.4)$$

donde $\underline{d}_t = \underline{c}_0 + \underline{c}_1 t$ con $\underline{c}_0, \underline{c}_1$ son vectores k -dimensionales reales. Por supuesto, la prueba de co-integración se enfoca en el rango de la matriz Π , la cual está relacionada con la matriz de correlación entre \underline{Z}_{t-1} y $\nabla \underline{Z}_t$. Veamos, asumiendo normalidad multivariada, Johansen usa la prueba basada en los estimadores de *ML* para detectar el rango Π . Para esto se usa análisis de correlación canónica. Específicamente, la idea consiste en mitigar el efecto de la parte estacionaria sobre la prueba de co-integración, es decir, se consideran dos regresiones lineales multivariadas preliminares.

$$\begin{aligned} \nabla \underline{Z}_t &= \sum_{i=1}^{p-1} \gamma_i \nabla \underline{Z}_{t-i} + \underline{d}_t + \underline{u}_t \\ \underline{Z}_{t-1} &= \sum_{i=1}^{p-1} \gamma_i^* \nabla \underline{Z}_{t-i} + \underline{d}_t^* + \underline{v}_t \end{aligned}$$

Sean $\hat{\underline{u}}_t$ y $\hat{\underline{v}}_t$ los resultados de mínimos cuadrados de las ecuaciones de regresión anteriores. Luego, se considera la regresión lineal multivariada:

$$\hat{\underline{u}}_t = \Pi \hat{\underline{v}}_t + \underline{\varepsilon}_t \quad (12.5)$$

Se puede verificar que las estimación de Π de la ecuación 12.5 y 12.4 son equivalentes. Es decir, bajo el supuesto de normalidad, se puede chequear el rango de Π probando los coeficiente de correlación canónica entre $\hat{\underline{u}}_t$ y $\hat{\underline{v}}_t$ ver sección 5.9.3 del libro [37] basados en las estadísticas cociente de verosimilitud ya que son pruebas anidadas sobre el rango de la matriz Π . Sean $\lambda_1^2 \geq \lambda_2^2 \geq \dots \geq \lambda_k^2$ las correlaciones canónicas al cuadrado ordenadas entre $\hat{\underline{u}}_t$ y $\hat{\underline{v}}_t$, las cuales son los valores propios al cuadrado de la matriz

$$\hat{\Sigma}_{11}^{-1} \hat{\Sigma}_{10} \hat{\Sigma}_{00}^{-1} \hat{\Sigma}_{01}$$

donde

$$\hat{\Sigma}_{00} = \frac{1}{T} \sum_{t=1}^T \hat{u}_t \hat{u}'_t, \quad \hat{\Sigma}_{11} = \frac{1}{T} \sum_{t=1}^T \hat{v}_t \hat{v}'_t, \quad \hat{\Sigma}_{01} = \frac{1}{T} \sum_{t=1}^T \hat{u}_t \hat{v}'_t$$

es decir

$$\hat{\Sigma}_{11}^{-1} \hat{\Sigma}_{10} \hat{\Sigma}_{00}^{-1} \hat{\Sigma}_{01} \underline{g}_i = \lambda_i \underline{g}_i.$$

Un primer estadístico considerado se basa en lo siguiente:

Para $m = 0, \dots, k-1$, considere la hipótesis nula $H(m) : \text{Rango}(\Pi) = m$
v.s. $H_a : \text{Rango}(\Pi) > m$. El estadístico es

$$L_{tr}(m) = -(T - kp) \sum_{i=m+1}^l \ln(1 - \lambda_i^2) \quad (12.6)$$

donde T, p, k con el tamaño de la muestra, el orden del VAR y la dimensión respectivamente. La idea es que bajo $H(m) : \text{Rango}(\Pi) = m$, los m correlaciones canónicas mas grandes son positivos, pero las restantes $k-m$ son cero. La distribución límite del estadístico $L_{tr}(m)$ no es una χ^2 , esta es una función del movimiento Browniano estándar, y los cuantiles deben ser encontrados via simulación.

El segundo estadístico es concerniente a la prueba

$$H_0 : \text{Rango}(\Pi) = m \quad v.s. \quad \text{Rango}(\Pi) = m + 1$$

$$L_{max}(m) = -(T - kp) \ln(1 - \lambda_{m+1}^2). \quad (12.7)$$

Al igual que la estadística $L_{tr}(m)$, la distribución límite de $L_{max}(m)$ también es un movimiento Browniano estándar.

Una vez el número de vectores de co-integración m es determinado, los vectores propios asociados con los m valores propios mas grandes $\{\lambda_i^2 | i = 1, \dots, m\}$ de el análisis de correlación canónica entre \hat{u}_t y \hat{v}_t pueden ser usado para estimar los vectores de co-integración.

Nota 12.6. Ver como se lleva a cabo la prueba de co-integración en modelos VARMA? En la exposición se explicará. Que pasa si procedemos a diferenciar las serie y continuar con el análisis para las series diferenciadas?

12.6 Estimación de Modelo de Corrección de Errores

Vamos a hablar acerca del modelo de corrección de errores asumiendo que el número de vectores de co-integración es conocido. Vamos a asumir que el proceso de errores $\{\underline{a}_t\}$ es Gaussiano. Con lo cual, resulta equivalente a usar la estimación vía Cuasi-máxima verosimilitud(QMLE), es decir, la estimación de máxima verosimilitud cuando la función de verosimilitud está mal especificada. En general, el estimador de máxima verosimilitud en presencia de errores de especificación no muestra las propiedades habituales. Los órdenes p, q puede ser encontrados usando criterios de información.

Modelos VAR

Se puede presentar dos casos para la estimación de los parámetros del ECM. El primer caso consiste en suponer que la matriz de los vectores de cointegración β es conocido y así el proceso co-integrado $\underline{W}_t = \beta' \underline{Z}_t$ está disponible, en este caso, el modelo se reduce:

$$\nabla \underline{Z}_t = \alpha \underline{W}_t + \underline{c}(t) + \sum_{i=1}^{p-1} \Phi_i^* \nabla \underline{Z}_{t-i} + \underline{a}_t, \quad (12.8)$$

el cual puede estimarse por el método de mínimos cuadrados ordinarios, y por lo tanto se puede usar la distribución normal de forma asintótica para hacer inferencia.

El otro caso consiste, en que β es desconocido. Con lo cual el modelo queda

$$\nabla \underline{Z}_t = \alpha \beta' \underline{Z}_{t-1} + \sum_{i=1}^{p-1} \Phi_i^* \nabla \underline{Z}_{t-i} + \underline{d}_t + \underline{a}_t \quad (12.9)$$

el cual involucra el producto de los parámetros y así requiere una estimación del tipo no-lineal. Aquí se usará la parametrización con $\beta' = [I_m : \beta_1']$ donde β_1 es una matriz $(k - m) \times m$. Note que la idea de usar esa parametrización es que la matriz de cointegración β es de rango completo m y satisface $\alpha \beta' = (\alpha P)(P \beta)'$ y sea identificable. Aquí se usa la estimación basada en QMLE y la distribución límite involucra funciones del movimiento Browniano.

12.7 Pronóstico de Procesos Integrados y Co-integrados

Si el objetivo es pronosticar, un modelo VAR es conveniente. Para un proceso VAR(P)

$$\underline{Y}_t = A_1 \underline{Y}_{t-1} + \cdots + A_p \underline{Y}_{t-p} u_t \quad (12.10)$$

el optimo pronóstico $h-$ pasos adelante con mínimo MSE es dada por la esperanza condicional, incluso si $\det(I_K - A_1 z - \cdots - A_p z^p)$ tiene raíces sobre el círculo unitario. Note que la optimalidad del pronóstico no depende de la condición de estacionariedad, por lo tanto el pronóstico $h-$ pasos adelante tiene forma:

$$\underline{Y}_t(h) = A_1 \underline{Y}_t(h-1) + \cdots + A_p \underline{Y}_t(h-p)$$

donde $\underline{Y}_t(j) = \underline{Y}_{t+j}$

13

Modelo Factorial Dinámico

La idea de los modelos factoriales nacen con Spearman para explicar el concepto de inteligencia. En forma general se busca que factores no observables \underline{f} , expliquen las variables de interés Z . En el análisis multivariado de series de tiempo, la ventaja del análisis factorial(dinámico(DFM)), consiste en que los factores pueden representar la evolución de la dinámica de ciertas series de tiempo multivariadas con un poco número de parámetros.

13.1 DFM para Series Estacionarias

Considere que $\underline{z}_1, \underline{z}_2, \dots, \underline{z}_T$, T realizaciones de una serie de tiempo estacionaria k -dimensional con $\underline{z}_t = (z_{1t}, \dots, z_{kt})'$. Asumamos que la serie está centrada, es decir se le sustrae la media tal que $\sum_{t=1}^T z_{jt} =$ para cada subserie j . Además sea \mathbf{Z} la matriz de datos de dimensión $T \times k$, tal que la t -ésima fila de la matriz es \underline{z}_t' . Asumamos que la dinámica de la evolución de las series se puede explicar por dos componentes ortogonales:

- ✓ La primera componente es común a todas las series , y es la responsable de las autocorrelaciones y autocorrelaciones cruzadas entre las series.
- ✓ La segunda componente es el ruido, también llamado componente específico o componente idiosincrático, el cual toma en cuenta pequeñas dinámicas que son específicas de cada serie.

Con esto en mente, el modelo DFM se define como sigue:

$$\underline{Z}_t = \mathbf{P}\underline{f}_t + \underline{\eta}_t, \quad (13.1)$$

donde $\underline{f}_t = (f_{1t}, f_{2t}, \dots, f_{rt})'$ es el vector de factores comunes r -dimensional, $\mathbf{P} = [p_{ij}]$ es la matriz de carga $k \times r$ de rango completo r , y $\underline{n}_t = (n_{1t}, \dots, n_{kt})'$ es una serie idiosincrática. Se asume que las componentes \underline{f}_t y \underline{n}_t son independientes. Sea \mathbf{p}_i la i -ésima fila de la matriz de carga \mathbf{P} , el cual es un vector fila $1 \times r$, con esto tenemos

$$Z_{it} = \mathbf{p}_i \cdot \underline{f}_t + n_{it}. \quad (13.2)$$

La notación matricial para los valores T del modelo DFM queda

$$\mathbf{Z} = \mathbf{F}\mathbf{P}' + \mathbf{N} \quad (13.3)$$

Donde \mathbf{F} la matriz $T \times r$ la matriz de factores donde la t -ésima fila es \underline{f}_t' , y \mathbf{N} la matriz $T \times k$, mas específicamente

$$\begin{bmatrix} z_{11} & z_{21} & \cdots & z_{k1} \\ z_{12} & z_{22} & \cdots & z_{k2} \\ \vdots & \vdots & \vdots & \vdots \\ z_{1T} & z_{2T} & \cdots & z_{kT} \end{bmatrix} = \begin{bmatrix} f_{11} & f_{21} & \cdots & f_{r1} \\ f_{12} & f_{22} & \cdots & z_{r2} \\ \vdots & \vdots & \vdots & \vdots \\ f_{1T} & f_{2T} & \cdots & f_{rT} \end{bmatrix} + \begin{bmatrix} p_{11} & p_{21} & \cdots & p_{k1} \\ p_{12} & p_{22} & \cdots & p_{k2} \\ \vdots & \vdots & \vdots & \vdots \\ p_{1r} & p_{2r} & \cdots & f_{kr} \end{bmatrix} + \begin{bmatrix} n_{11} & n_{21} & \cdots & n_{k1} \\ n_{12} & n_{22} & \cdots & n_{k2} \\ \vdots & \vdots & \vdots & \vdots \\ n_{1T} & n_{2T} & \cdots & n_{kT} \end{bmatrix}$$

Sea $\underline{c}_t = \mathbf{P} \underline{f}_t = (c_{1t}, \dots, c_{kt})'$ el vector $k \times 1$ representa el componente común de cada serie y es generado por los r factores comunes no observables en \underline{f}_t , donde para efectos prácticos $r \ll k$. Note mas claramente que

$$\begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1r} \\ p_{21} & p_{22} & \cdots & p_{2r} \\ \vdots & \vdots & \vdots & \vdots \\ p_{k1} & f_{k2} & \cdots & f_{kr} \end{bmatrix} \begin{bmatrix} f_{1t} \\ f_{2t} \\ \vdots \\ f_{rt} \end{bmatrix} = \begin{bmatrix} p_{11}f_{1t} + p_{12}f_{2t} + \cdots + p_{1r}f_{rt} \\ p_{21}f_{1t} + p_{22}f_{2t} + \cdots + p_{2r}f_{rt} \\ \vdots \\ p_{k1}f_{1t} + p_{k2}f_{2t} + \cdots + p_{kr}f_{rt} \end{bmatrix} = \begin{bmatrix} c_{1t} \\ c_{2t} \\ \vdots \\ c_{kt} \end{bmatrix}.$$

De manera análoga al análisis factorial clásico, hay que imponer condiciones sobre las cargas y los factores para poder identificar el modelo factorial en 13.1. Esto es porque para una matriz \mathbf{Q} invertible $r \times r$, se tiene

$$\underline{c}_t = \mathbf{P} \underline{f}_t = \underline{c}_t = \underbrace{\mathbf{P} \mathbf{Q}^{-1}}_{\mathbf{P}^*} \underbrace{\mathbf{Q} \underline{f}_t}_{\underline{f}_t^*} = \mathbf{P}^* \underline{f}_t^*$$

Vale la pena decir que se pueden escoger dos enfoques para imponer las restricciones sobre \mathbf{P} y \underline{f}_t :

1. Requerir que $\Gamma_f(0) = E[\underline{f}_t \underline{f}_t'] = I_r$, lo cual implica que

$$\mathbf{Q} \mathbf{Q}' = I$$

y así que \mathbf{Q} debe ser una matriz ortonormal.

2. Otro requisito puede ser que $\mathbf{P}'\mathbf{P} = I$, lo cual implica la misma condición anterior

$$\mathbf{Q}\mathbf{Q}' = I.$$

Es decir, ambos requisitos llegan a la misma suposición acerca de la matriz \mathbf{Q} . Se va entonces a asumir inicialmente que $\mathbf{P}'\mathbf{P} = I$, al igual que $\Gamma_f(0)$ sea una matriz diagonal. Lo cual implica que el DFM es identificable ante rotaciones. Con lo cual, una vez el modelo sea haya ajustado a los datos, siempre es posible explorar una rotación en los factores con el objeto de que sea un modelo fácil de entender. Lo cual está relacionado con la ortonormalidad de matriz de carga.

Adicionalmente, vamos a asumir que los factores pueden aproximarse bien por medio de un modelo VARMA causal y reversible, es decir

$$\Phi(B)\underline{f}_t = \Theta(B)\underline{a}_t, \quad (13.4)$$

donde el proceso de ruido r -dimensional $\{\underline{a}_t\}$, es un ruido blanco con Σ_a una matriz de varianzas y covarianzas diagonal. Además se asume que los factores están no autocorrelacionados con el proceso de ruido idiosincrático \underline{n}_t , es decir, $E[\underline{a}_t \underline{n}'_{t-h}] = 0$, para todo rezago h .

En la práctica, es usual suponer dos tipos de condiciones sobre ruido idiosincrático \underline{n}_t :

- ✓ $\{\underline{n}_t = \underline{e}_t\}$ es una sucesión independiente e idénticamente distribuida de vectores aleatorios de media cero y matriz de covarianza Σ_e , además que los factores no estén correlacionados con este proceso, es decir, $E[\underline{a}_t \underline{e}'_{t-h}] = 0$. A este modelo DFM se le llama modelo DFM exacto o EDFM, así es el modelo es identificable excepto por rotaciones del factor.
- ✓ Otro supuesto que es común proponer, está en el contexto de alta dimensionalidad, es decir cuando k es muy grande. En este caso se permite algún tipo de correlación o de correlación cruzada sobre $\{\underline{n}_t\}$. El modelo es identificable cuando ambos T, k tienden a infinito y así el modelo es llamado un modelo DFM aproximado (ADFM).

Propiedades de las Matrices de Covarianza

Para una serie de tiempo estacionaria, la información disponible para estimar al DFM está en las matrices de covarianza. Sean $\Gamma_f(h)$ y $\Gamma_n(h)$ las

matrices de autocovarianzas en el rezago h del proceso de factores y del ruidos respectivamente. Entonces tenemos que:

$$\Gamma_z(h) = \mathbf{P}\Gamma_f(h)\mathbf{P}' + \Sigma_e(h), \quad h \geq 0. \quad (13.5)$$

el análisis factorial dinámico depende de la ecuación anterior.

Nota 13.1. *Tal cual sucede en el análisis factorial estático, si se hace una estandarización(cambio de escala) del proceso $\{\underline{Z}_t\}$, los resultados pueden diferir si no hay un cambio de escala. Las variables con varianzas influyen mucho el análisis. Por lo que se recomienda trabajar con los datos estandarizados y sin estandarizar y comparar los resultados.*

El DFM Exacto

Consideremos el caso cuando k es finito y $\{\underline{n}_t = \underline{e}_t\}$, es decir, es un ruido blanco tal que para la ecuación 13.5 obtenemos

$$\Gamma_z(0) = \mathbf{P}\Gamma_f(0)\mathbf{P}' + \Sigma_e, \quad (13.6)$$

$$\Gamma_z(h) = \mathbf{P}\Gamma_f(h)\mathbf{P}', \quad h > 0. \quad (13.7)$$

note que el rango de la matriz $\Gamma_z(h)$ es el mismo que el de $\Gamma_f(h)$ ya que \mathbf{P} tiene rango completo. Además el número de factores viene dado por $r = \max_h\{rango[\Gamma_z(h)]|h > 0\}$. Así que comprobar los rangos de las matrices de autocovarianza muestral de \underline{Z}_t , nos da información acerca de el número de factores comunes.

Nota 13.2. *Dado que toda la dinámica de las series está dirigida por los factores comunes f_t , vamos a asumir de forma natural que ninguna combinación de los factores comunes es ruido blanco. Además si cada unos de los factores latentes f_t de son independientes entonces $\Gamma_f(h)$ es una matriz diagonal. Además se puede verificar que $\Gamma_z(h)$ es simétrica para todo h .*

Estas características se usarán para verificar la existencia de un DFM con factores comunes independientes. Vamos a suponer que la estructura de la matriz de covarianza Σ_e de la componente específica n_t es de las siguientes dos formas:

Caso I. La matriz de covarianza es $\Sigma_e = \sigma^2 I$, por lo tanto podemos ver al pos-multiplicar por la j -ésima columna de \mathbf{P} a $\Gamma_z(0)$ tenemos

$$\Gamma_z(0)\mathbf{p}_{.j} = (\gamma_{f,j}^2 + \sigma^2)\mathbf{p}_{.j}$$

con $\gamma_{f,j}^2$ siendo la varianza del j -ésimo factor común y $1 \leq j \leq r$. Es decir, $\mathbf{p}_{.j}$ es el j -ésimo vector propio de $\Gamma_z(0)$ con el valor propio asociado $\gamma_{f,j}^2 + \sigma^2$.

Por otro lado, asuma que $\mathbf{P}_\perp(k \times (k - r))$ es la matriz complemento ortogonal de \mathbf{P} en \mathbb{R}^k , tal que $(\mathbf{P}_\perp)' \mathbf{P} = 0$ una matriz de ceros $((k - r) \times r)$. Sea $= \mathbf{p}_{.j}^*$ la j -ésima columna de \mathbf{P}_\perp , de manera análoga tenemos

$$\Gamma_z(0) \mathbf{p}_{.j}^* = \sigma^2 \mathbf{p}_{.j}^*$$

es decir, $\mathbf{p}_{.j}^*$ es también un vector propio de $\Gamma_z(0)$ asociado con el valor propio σ^2 . En general podemos ver que los r valores propios de $\Gamma_z(0)$ asociados con los r valores propios mas grandes, forman la matriz de cargas de modelo exacto DFM.

Caso II. La matriz de covarianza $\Sigma_e = \text{Diag}\{\sigma_1^2 - \sigma^2, \dots, \sigma_k^2 - \sigma^2\}$. Sea $\sigma^2 = \max_{1 \leq j \leq k} \{\sigma_j^2\}$ la varianza mas grande de los ruidos. Entonces tenemos que

$$\Sigma_e = \sigma^2 I + D_0,$$

donde $D_0 = \text{Diag}\{\sigma_1^2 - \sigma^2, \dots, \sigma_k^2 - \sigma^2\}$. Asumamos ahora que $\gamma_{f,r}^2$ es la varianza mas pequeña de los factores comunes f_{jt} , para $1 \leq j \leq r$. Llevando a cabo el mismo procedimiento anterior, tenemos

$$\Gamma_z(0) \mathbf{p}_{.r} = (\gamma_{f,r}^2 + \sigma^2)(\mathbf{p}_{.r} + \delta_t),$$

donde $\delta_r = D_0 \mathbf{p}_{.r} / (\gamma_{f,r}^2 + \sigma^2)$. En este caso $\mathbf{p}_{.r}$ no es exactamente un vector propio ya que en general δ_r no es el vector cero en general. Sin embargo si, $\gamma_{f,r}^2$ es mucho mas grande que σ^2 , entonces δ_r debería ser pequeño o incluso se puede despreciar, lo cual en este caso sería aproximadamente un vector propio de $\Gamma_z(0)$ con valor propio $\gamma_{f,r}^2 + \sigma^2$. En resumen, si el cociente $\frac{\sigma^2}{\gamma_{f,r}^2}$ (cociente ruido a señal) es muy pequeño, entonces se cumplirían las condiciones de el caso I para el casi II. Es decir, el modelo debería ser mas útil y fácil de identificar si el cociente ruido a señal es pequeño.

El DFM Aproximado

En este caso, el modelo no es identificable en muestras finitas, pero puede ser identificable cuando ambos k, T tiende a infinito bajo ciertas condiciones. Cuando η_t tiene dinámica de dependencia, entonces 13.7 resulta en

$$\Gamma_z(h) = \mathbf{P}\Gamma_f(h)\mathbf{P}' + \Gamma_n(h), \quad h > 0, \quad (13.8)$$

asumiendo que \underline{f}_t no está correlacionado con $\underline{\eta}_{t-h}$. con lo cual, el rango de $\Gamma_z(h)$ no da información acerca de el número de factores comunes r . Sin embargo, bajo condiciones similares como las de caso II para el el modelo EDFM, se puede identificar la matriz de pesos. Esto se alcanza si

$$\lim_{k \rightarrow \infty} \frac{1}{k^2} \sum_{i=1}^k \sum_{j=1}^k |\gamma_{n,ij}| \rightarrow 0,$$

donde $\gamma_{n,ij}$ es la covarianza de n_{it} y n_{jt} .

Ejemplo 13.3. Vamos a demostrar como las matrices de autocovarianza muestran las características de los DFM. Vamos a considerar las tasas de crecimiento del producto interno bruto estandarizado de 19 países de la zona Euro.

13.2 Factores Dinámicos y Modelos VARMA

De la ecuación del modelo exacto, con $\underline{\eta}_t$ siendo ruido blanco se puede implicar que $k - r$ combinaciones lineales de las series son ruido blanco y r combinaciones lineales contienen todas las dinámicas. Sea $\mathbf{P}_\perp(k \times (k - r))$ la matriz complemento ortogonal de la matriz \mathbf{P} . Entonces, las $k - r$ combinaciones lineales $\underline{Y}_{1t} = (\mathbf{P}_\perp)' \underline{Z}_t$ son ruido blanco y las r combinaciones lineales $\underline{Y}_{2t} = \mathbf{P}' \underline{Z}_t$ contiene los factores comunes. Sea ha probado que lo contrario también es cierto, es decir, consideremos un un proceso $VAR(p)$ para el proceso $\{\underline{Z}_t\}$

$$\underline{Z}_t = \sum_{i=1}^p \Phi_i \underline{Z}_{t-i} + \underline{a}_t.$$

si fuera posible encontrar una matriz $G(k \times k)$ tal que $\underline{Y}_t = G \underline{Z}_t$ satisface el modelo

$$\underline{Y}_t = \sum_{i=1}^p \Phi_i^* \underline{Y}_{t-i} + \underline{a}_t^*$$

tal que $\Phi_i^* = G \Phi_i G^{-1}$ y $\Sigma_a^* = G \Sigma_a G'$, y la nueva serie \underline{Y}_t puede partitionarse como $\underline{Y}_t = (\underline{Y}'_{1t}, \underline{Y}'_{2t})'$ donde \underline{Y}'_{1t} es un proceso $k - r$ dimensional de ruido blanco y \underline{Y}'_{2t} sigue un proceso $VAR(P)$, entonces $\{\underline{Z}_t\}$ sigue un DFM.

De hecho, se puede verificar que si \underline{Z}_t sigue un DFM con factores que siguen un proceso $VARMA(p_1, q_1)$, entonces la serie \underline{Z}_t sigue proceso $VARMA(p_2, q_2)$ con $p_1 = p_2$ y $q_2 = \max(p_1, q_1)$, pero este proceso es no identificable.

13.3 Ajuste de un DFM estacionario a los datos

Sean $\{\underline{Z}_1, \dots, \underline{Z}_T\}$ los datos que está ajustado por la media, y $\mathbf{Z}(T \times k)$ la matriz de datos, donde $\mathbf{Z}' = [\underline{Z}_1, \dots, \underline{Z}_T]$. Con esto, los estimadores de las matrices de autocovarianza son

$$\hat{\Gamma}_z(h) = \sum_{t=h+1}^T \underline{Z}_t \underline{Z}'_{t-h} / T,$$

y en particular $\hat{\Gamma}_z(0) = \mathbf{Z}' \mathbf{Z} / T$. La idea ahora consiste en explicar procedimiento para estimar la matriz de cargas $P(k \times r)$, la matriz $\mathbf{F}(T \times r)$ de los factores comunes(predecir), y los parámetros de los modelos para los factores y los ruidos. Inicialmente vamos a suponer que el número de factores r es conocido, luego procederemos a discutir como se selecciona cuando es desconocido. Asumiremos que $T < k$.

Estimación Usando Componentes Principales(PC)

Bajo ciertas condiciones, se puede obtener una estimación consistente para ambos, EDFM y ADFM vía el análisis de valores-vectores propios de la matriz de covarianza muestral $\hat{\Gamma}_z(0)$. Con r dado, la matriz de cargas \mathbf{P} es estimada por los vectores propios normalizados de $\hat{\Gamma}_z(0)$ correspondiente a los r valores propios mas grandes. Sea $\hat{\mathbf{P}}$ la estimación de la matriz de cargas, y puesto que $\hat{\Gamma}_z(0)$ es simétrica y definida positiva, para $T > k$, tenemos que $\hat{\mathbf{P}}' \hat{\mathbf{P}} = I$. Los factores latentes son estimados por $\hat{f}_t = \hat{\mathbf{P}}' \underline{Z}_t$, los cuales son los primeros r componentes principales de la matriz de datos \mathbf{Z} . Ver libro [30] página 302 para detalles de como se encuentra la estimación minimizando la suma de cuadrados $MSE(\mathbf{F}, \mathbf{P}) = \frac{1}{kT} \sum_{t=1}^T (\underline{Z}_t - \hat{\mathbf{P}} \hat{f}_t)' (\underline{Z}_t - \hat{\mathbf{P}} \hat{f}_t)$. Stock y Watson, Bai y Ng mostraron que la estimación mediante PC es consistente cuando $k, T \rightarrow \infty$ bajo condiciones débiles de las autocorrelaciones y de las correlaciones cruzadas de los ruidos.

Una vez los factores son estimados, los componentes idiosincráticos se pueden computar por medio de $\hat{n}_t = (I - \hat{\mathbf{P}} \hat{\mathbf{P}}') \underline{Z}_t$, y podemos estimar los modelos escalares $ARMA$ para cada componente n_{it} . Sea $\hat{N}' = [\hat{n}_1, \dots, \hat{n}_T]$,

la matriz de los ruidos estimados, ahora la estimación de la matriz de covarianza de los ruidos es $\hat{\Gamma}_n(0) = \hat{N}'\hat{N}/T$, la cual es singular. Sin embargo, los elementos de la diagonal pueden estimarse por $\sum_{t=1}^T \hat{n}_{it}^2/T$.

Los modelos para los factores pueden también ser obtenidos por el análisis univariado de cada serie de factores \hat{f}_{it} .

El Estimador Combinado PC

Para el EDFM, una estimación mas eficiente que el PCA, sugerido por Peña y Box, se puede obtener combinando la información de las matrices de autocovarianza, mas específicamente

$$C = \sum_{i=1}^{h_0} \hat{\Gamma}_z(i),$$

se puede estimar \mathbf{P} combinando los r vectores propios enlazados con los valores propios mas grandes de la matriz C . Sin embargo, como las matrices $\hat{\Gamma}_z(h)$ no son simétricas, ellas pueden tener valores propios complejos. Una mejor alternativa, es usar una matriz simétrica definida como sigue

$$\mathbf{L} = \sum_{i=1}^{h_0} \hat{\Gamma}_z(i)\hat{\Gamma}_z(i)'$$

y con esta matriz hacer el PCA con los r valores propios mas grandes.

Estimador PC Generalizado

En el caso de heterogeneidad, las varianzas de los ruidos componentes son diferentes, y se puede usar el método de mínimos cuadrados generalizados para mejorar la eficiencia de la estimación. La función objetivo es

$$MSE(\mathbf{F}, \mathbf{P}) = \frac{1}{kT} \sum_{t=1}^T (\mathcal{Z}_t - \mathbf{P}\hat{f}_t)' \Sigma_n^{-1} (\mathcal{Z}_t - \mathbf{P}\hat{f}_t).$$

Como Σ_n^{-1} es desconocida, la minimización de llevarse a cabo en dos etapas. Primero se hace la estimación usual como se explicó anteriormente, y se obtiene $\hat{n}_t = (I - \hat{\mathbf{P}}\hat{\mathbf{P}}')\mathcal{Z}_t$, los los cuales se puede estimar la matriz de varianza y covarianza Σ_n , pero ajustándola debido a que esta es singular.

Estimación vía Máxima Verosimilitud

Hay dos enfoques para la estimación de máxima verosimilitud, en donde se asume inicialmente que el proceso $\{\mathcal{Z}_t\}$ es Gaussiano. El primero es

una extensión directa de los modelos factoriales clásicos(estáticos), es decir se asume que \underline{f}_t es un vector aleatorio pero sin dinámica temporal, y así $Z_t \sim N_k(\mathbf{0}, \Gamma_z(0))$ on $\Gamma_z(0) = \mathbf{P}\Gamma_f(0)\mathbf{P}' + \Sigma_e$. En este caso los parámetros son \mathbf{P} , los r elementos de diagonal $\Gamma_f(0)$, y los k elementos de la diagonal de Σ_e . En este caso la serie Z_t es independiente de sus retardos y la función de verosimilitud queda

$$L(\mathbf{P}, \Gamma_f(0), \Sigma_e) = -\frac{T}{2} \log |\Gamma_z(0)| - \frac{T}{2} \text{tr}[\hat{\Gamma}_z(0)\hat{\Gamma}_z^{-1}(0)].$$

El segundo enfoque es considerar que \underline{f}_t como parámetros o variables aleatorias con dependencia temporal, y usar la distribución condicional de Z_t como $N_k(\mathbf{P}\underline{f}_t, \Sigma_e)$, y así la función de verosimilitud queda

$$L(\mathbf{P}, \Gamma_f(0), \Sigma_e) = -\frac{T}{2} \log |\Sigma_e| - \frac{1}{2} \sum_{t=1}^T (Z_t - \mathbf{P}\underline{f}_t)' \Sigma_e^{-1} (Z_t - \mathbf{P}\underline{f}_t),$$

siendo los parámetros \mathbf{P} , \mathbf{F} , y los k elementos de la diagonal de Σ_e . La minimización de 13.3 se lleva a cabo poniendo el modelo como un modelo de espacio-estado y computar la verosimilitud usando el filtro de Kalman. Ver libro [30] para la implementación del algoritmo para minimización.

Selección del Número de Factores

Varias pruebas han sido propuestas en la literatura para seleccionar el número de factores. Hay pruebas basadas en comprobar el rango de las matrices de autocovarianza. En un modelo EDFM, el máximo rango debería ser r y este es alcanzado por la suma de algunas matrices de autocovarianzas con $h > 0$. Estas pruebas son útiles cuando hay fuerte autocorrelación en los ruidos como en el modelo ADFM, ya que las matrices de autocovarianza pueden ser rango completo. Estas pruebas trabajan bien cuando $T >> k$.

Otro tipo de pruebas usan la idea de separar los valores propios grandes de las matrices de covarianzas de los pequeños, y seleccionar r como el número de valores propios grandes. Un tercer tipo de pruebas consiste en usar criterios de información.

Prueba del Rango vía Correlación Canónica

La idea es chequear el rango de algunas matrices de momentos, usando el análisis de correlación canónica y el estadístico χ^2 . Para $h > 0$, existe una matriz $\mathbf{P}_\perp(k \times (k - r))$ tal que $\Gamma_z(h)\mathbf{P}_\perp = \mathbf{0}$. Con esto, se puede verificar

que las series $\mathbf{P}'_{\perp} \mathcal{Z}_t$ no tienen correlación ni correlación cruzada para todos los rezagos, y también no correlacionados con $\mathbf{P}'_{\perp} \mathcal{Z}_{t-h}$. Consideremos la matriz($k \times k$) de correlación canónica

$$\mathbf{M}(h) = [E(\mathcal{Z}_t \mathcal{Z}'_t)]^{-1} E(\mathcal{Z}_t \mathcal{Z}'_{t-h} [E(\mathcal{Z}_{t-h} \mathcal{Z}'_{t-h})]^{-1} E(\mathcal{Z}_{t-h} \mathcal{Z}'_t)),$$

asumiendo por simplicidad que $E[\mathcal{Z}_t] = 0$. Asumiendo que el EDFM satisface que $\Gamma_z(h) = \Gamma_z(-h)$, se tiene que

$$\mathbf{M}(h) = [\Gamma_z^{-1}(0)\Gamma_z(h)]^2.$$

Por ahora, asumamos que $rango[\Gamma_f(r)] = r$, el cual también es el rango de $\Gamma_z(h)$, por lo tanto $rango[\mathbf{M}(h)] = r$. el número de correlaciones canónicas cero entre \mathcal{Z}_{t-h} y \mathcal{Z}_t es dado por número de valores propios cero de la matriz $\mathbf{M}(h)$, el cual es $k - r$. Con lo cual el número de factores comunes r es equivalente al número de correlaciones canónicas diferentes de cero entre \mathcal{Z}_{t-h} y \mathcal{Z}_t y así depende de los valores propios de la matriz $M(h)$.

Para llevar a cabo la prueba del número de factores, se procede de forma secuencial como lo propuso Peña y Poncela(2008). Para esto, considere los valores propios ordenados $\hat{\lambda}_1 \geq \hat{\lambda}_2 \dots \geq \hat{\lambda}_k$ de la matriz $\hat{\mathbf{M}}(h)$. Si hay r factores, los valores propios $\hat{\lambda}_{r+1}, \dots, \hat{\lambda}_k$ son estimaciones de correlaciones cuadradas iguales a cero, las cuales tienen varianza asintótica $1/(T - h)$. Con lo cual, el estadístico $-(T - h) \log(1 - \hat{\lambda}_j) \simeq (T - h)\hat{\lambda}_j$. Esto permite demostrar que

$$S_{h-r} = -(T - h) \sum_{j=r+1}^k \log(1 - \hat{\lambda}_j) \quad (13.9)$$

es asintóticamente una $\chi^2_{(k-r)^2}$. Esta prueba se aplica secuencialmente. Es decir, se empieza con $r = 0$, y si la hipótesis es rechazada, el valor $r = 1$ es probado, y así sucesivamente. El procedimiento se detiene cuando la hipótesis de r factores, o r correlaciones canónicas diferente de cero, no puede ser rechazada.

Esta prueba trabaja bien cuando T/k es grande. También puede suceder que la prueba falle cuando matrices de covarianza para ciertos rezagos no son de rango completo.

Existe otra forma de encontrar r , y es mediante el criterio de saltos en los valores propios y se basa en los cocientes de los valores propios consecutivos de la matriz de covarianza o de la suma de las matrices de covarianza. Aquí, se procede a computar los valores propios $\lambda_1 \geq \lambda_2 \dots \geq \lambda_k$ de la matriz combinada

$$\mathbf{M} = \sum_{h=1}^{h_0} \hat{\Gamma}_z(h) \hat{\Gamma}_z(h)', \quad (13.10)$$

con h_0 un entero positivo pre-especificado. Se selecciona r como

$$\hat{r} = \arg \min_{1 \leq i \leq r^*} \frac{\lambda_{i+1}}{\lambda_i},$$

para algún $r^* = \alpha k$, donde $0 < \alpha < 1$, por ejemplo $\alpha = 0.2$.

Criterios de Información

Similar como se ha hecho en otros escenarios, los criterios de información pueden usarse para identificar el número de factores en un modelo factorial. Tenemos por ejemplo el criterio *BIC* defino como:

$$BIC(t) = \log \hat{\sigma}_r^2 + r \frac{\log T}{T}.$$

Otro criterio es como sigue:

$$BNG(r) = \log \hat{\sigma}_r^2 + r \frac{\log \min(T, k)}{\min(T, k)},$$

donde

$$\sigma_r^2 = MSE(\hat{\mathbf{P}}_r, \hat{\mathbf{F}}_r) = \frac{1}{kT} \sum_{t=1}^T (\tilde{z}_t - \hat{\mathbf{P}}_r \hat{f}_t^r)' (\tilde{z}_t - \hat{\mathbf{P}}_r \hat{f}_t^r).$$

Pronósticos con Modelos Factoriales Dinámicos

El pronóstico con DFM se realiza ajustando modelos escalares a cada factor y cada serie de ruido, y una vez hecho esto se generan los pronósticos de los modelos modelos ajustados, digamos $\hat{f}_t(h)$ y $\hat{n}_t(h)$, donde h es el horizonte de pronostico en el origen t . Entonces

$$\hat{z}_t(h) = \hat{\mathbf{P}} \hat{f}_t(h) + \hat{n}_t(h).$$

Ejemplo 13.4. Como ejemplo ilustrativo, consideremos las tasas de crecimiento del PIB de los 19 países de la zona Euro usados anteriormente.

13.4 Componentes Principales

13.4.1 PCA Estático

El PCA(estático) o para datos independientes, se ha mostrado que es útil también para series de tiempo. Hay algunas diferencias en aplicar directamente el PCA a datos de series de tiempo. La principal es que este se basa

en la matriz de covarianza($\Gamma(0)$), y no usa las autocovarianzas en otros otros retardos. Lo cual hace que las distribuciones límite de los valores y vectores propios sean diferentes para series de tiempo. Después veremos el DPCA, el cual es un procedimiento mas general.

Hay al menos dos formas de introducir el PCA. Una forma es la descomposición de varianza, y la otra es a través de la óptima reconstrucción o interpolación. Vamos a iniciar con la descomposición de varianza. Sea $\underline{Z} = (Z_1, \dots, Z_k)'$ un vector aleatorio k -dimensional de media \underline{Q} y matriz de covarianza definida positiva $\Gamma(0)$. La primera componente principal(PC) Y_1 de \underline{Z} es una combinación lineal $Y_1 = \underline{c}_1' \underline{Z}$, tal que $\underline{c}_1' \underline{c}_1 = 1$ y $Var(Y_1)$ alcanza el máximo entre todas las posibles combinaciones lineales de \underline{Z} . La segunda PC de \underline{Z} es una combinación lineal de \underline{Z} , llamémosla $Y_2 = \underline{c}_2' \underline{Z}$, tal que tiene la segunda varianza mas grande entre todas las combinaciones lineales de \underline{Z} y satisface $\underline{c}_2' \underline{c}_2 = 1$ y $\underline{c}_2' \underline{c}_1 = 0$ (e.d., Y_1, Y_2 son no correlacionadas). En forma general, la i -ésima PC de \underline{Z} es una combinación lineal de \underline{Z} , llamémosla $Y_i = \underline{c}_i' \underline{Z}$, tal que tiene la i -ésima varianza mas grande entre todas las combinaciones lineales de \underline{Z} y satisface $\underline{c}_i' \underline{c}_i = 1$ y $\underline{c}_i' \underline{c}_j = 0$ para $j = i - 1, \dots, 1$.

Nota 13.5. *Recuerde que la condición de la normalización $\underline{c}_i' \underline{c}_i = 1$ se requiere para controlar el efecto de escalamiento, de otra forma el máximo no tiene sentido.*

Ahora el PCA para un proceso estacionario $\{\underline{Z}_t\}$ con matriz de varianza-covarianza definida positiva $\Gamma(0)$ puede obtenerse de la descomposición espectral de $\Gamma(0)$. Mas precisamente, sea $(\lambda_i, \underline{e}_i)$ el i -ésimo valor propio-vector propio de $\Gamma(0)$, tal que se satisface $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_k$. La descomposición espectral es tal que:

$$\Gamma(0) = \lambda_1 \underline{e}_1' \underline{e}_1 + \dots + \lambda_k \underline{e}_k' \underline{e}_k.$$

Entonces, la i -ésima PC es $Y_{it} = \underline{e}_i' \underline{Z}_t$. De la definición, $Var(Y_{it}) = Var(\underline{e}_i' \underline{Z}_t) = \underline{e}_i' \Gamma(0) \underline{e}_i = \lambda_i$. Ahora, puesto que $tr[\Gamma(0)] = \sum_{i=1}^k \sigma_{ii} = \sum_{i=1}^k \lambda_i$, la proporción de variabilidad de \underline{Z}_t que es explicada por la i -ésima componente principal es $\lambda_i / (\sum_{i=1}^k \lambda_i)$. De forma similar, la proporción de variabilidad explicada por las primeras m componentes principales es $(\sum_{i=1}^m \lambda_i) / (\sum_{i=1}^k \lambda_i)$. Idealmente, si un cantidad pequeña m de componentes principales explican una proporción alta de variabilidad, entonces se puede enfocar el análisis basados en las m primeras componentes principales.

Ahora, con datos, supongamos que tenemos una muestra $\{\underline{z}_1, \dots, \underline{z}_T\}$ del proceso estocástico estacionario $\{\underline{Z}_t\}$. Por simpleza, asumamos que el proceso es ajustado por la media, es decir $\sum_{t=1}^T \underline{z}_t / T = \underline{0}$. Definamos la matriz $Z = [z_{(1)}, \dots, z_{(k)}]$, donde $z_{(j)} = (z_{j1}, \dots, z_{jT})'$ es el vector de observaciones para la j -ésima componente z_{jt} . Entonces la matriz de covarianzas de \underline{Z}_t es estimada por la matriz de covarianza muestral

$$\hat{\Gamma}(0) = Z'Z/T$$

y los pares de valor y vector propio de esta matriz, $(\lambda_i, \underline{c}_i)$, pueden usarse para construir las componentes principales. La primera componente principal es dada $\underline{y}_1 = Z\underline{c}_1$, con

$$\frac{1}{T} Z' Z \underline{c}_1 = \lambda_1 \underline{c}_1, \quad \underline{c}_1' \underline{c}_1 = 1.$$

Ahora, note que la pre-multiplicar la ecuación anterior por Z , tenemos

$$\frac{1}{T} Z Z' \underline{y}_1 = \lambda_1 \underline{y}_1,$$

es decir, la primera pc \underline{y}_1 de \underline{Z}_t es proporcional al valor propio correspondiente al valor propio de la matriz ZZ' . El otro enfoque para deducir las componentes principales es a través de la óptima reconstrucción o interpolación de los datos, ver [30] página 25.

13.4.2 Propiedades de las PC

Las componentes principales $\underline{y}_t = (y_{1t}, \dots, y_{kt})'$ de un proceso estocástico estacionario \underline{Z}_t son transformaciones lineales de los datos observados, y son ordenadas de acuerdo a su variabilidades tal que y_{1t} tiene la varianza mas grande. En particular $Var(y_{it}) = \lambda_i$, el cual es el i -ésimo valor propio mas grande de la matriz de covarianza de \underline{Z}_t . Ahora, como $\sum_{i=1}^k Var(z_{it}) = \sum_{i=1}^k \gamma_{ii}(0) = tr[\Gamma(0)]$, y también note que $tr[\Gamma(0)] = \sum_{i=1}^k \lambda_i$, es decir, las pc retienen la variabilidad total del proceso \underline{Z}_t .

Ahora, sea \underline{e}_i el valor propio asociado con el valor propio λ_i de $\Gamma(0)$, tal que tiene norma 1. Tenemos entonces que $\underline{y}_t = P' \underline{Z}_t$, donde $P = [\underline{e}_1, \dots, \underline{e}_k]$ es la matriz de los vectores propios. Esta matriz es ortonormal, tal que $PP' = P'P = I$. Por lo tanto $1 = |PP'| = |P||P'|$. Así, podemos ver que $Cov(\underline{y}_t) = P'Cov(\underline{Z}_t)P$, con lo cual podemos ver que $|Cov(\underline{y}_t)| = |Cov(\underline{Z}_t)|$, es decir, la varianza generalizada de \underline{y}_t es la misma que la del proceso \underline{Z}_t .

Note que de la misma forma que se hace con componentes principales para datos iid, también podemos reducir la dimensión del proceso vectorial,

es decir, escogiendo un número menor de componentes principales tal que la proporción de variabilidad en \mathcal{Z}_t explicada por las primeras m componentes principales sea $(\sum_{i=1}^m \lambda_i) / (\sum_{i=1}^k \lambda_i)$.

Ejemplo 13.6. Usaremos los índices de precios al consumidor mensuales de 33 países Europeos desde enero de 2000 hasta octubre de 2015. Estos datos están en el paquete SLBDD y se puede acceder por medio de data(CPIEurope200015).

Nota 13.7. Si bien, las componentes principales \underline{y}_t son contemporáneamente no correlacionadas, ellas pueden estar dinámicamente o serialmente correlacionadas, lo cual implica que debe analizarse de forma conjunta. en este caso se muestra que :

- a. Los valores propios de la matriz de covarianza muestral de un proceso estacionario y con estructura de autocorrelación vectorial están asintóticamente correlacionados.
- b. Los valores y vectores propios son asintóticamente dependientes.

13.5 Componentes Principales Dinámicas(PCA Dinámico(DPCA))

Vale la pena recordar que el PCA de la anterior sección sólo usa información contemporánea. La idea ahora consistirá en usar información de las variables retardadas, y es lo que conocemos como Componentes Principales Dinámicas(PCA Dinámico(DPCA)), es decir combinaciones lineales de las series usando todos los adelantos y retardos de los datos. Consideraremos el proceso k -dimensional de media cero $\{\mathcal{Z}_t | -\infty < t < \infty\}$. Definimos la primera componente principal dinámica DPC como la combinación lineal de todos los valores de la serie:

$$f_t = \sum_{h=-\infty}^{\infty} \underline{c}_h' \mathcal{Z}_{t-h}, \quad (13.11)$$

donde \underline{c}_h son vectores k -dimensionales tal que f_t da una reconstrucción óptima de los datos usando todos los adelantos y retardos, es decir, la primera DPC minimiza

$$E \left[\left(\mathcal{Z}_t - \sum_{j=-\infty}^{\infty} \underline{\beta}'_j f_{t+j} \right)' \left(\mathcal{Z}_t - \sum_{j=-\infty}^{\infty} \underline{\beta}'_j f_{t+j} \right) \right], \quad (13.12)$$

para algunos vectores $k \times 1$ β_j . Se demostró que los $\underline{\alpha}_k$ es la transformada de Fourier de las PCs de las matrices de cruzadas-espectral para cada frecuencia, y $\underline{\beta}_j$ es la transformada inversa de las conjugadas de las mismas PC. Por supuesto que el procedimiento debe ser adaptado para muestras finitas, incluyendo el número de retardos tanto en 13.11 y 13.12. Todo esto en el dominio de la frecuencia.

13.5.1 DPCs de un lado o cara

Hay una propuesta para resolver este problema en el dominio del tiempo, que es mas eficiente que el del dominio de la frecuencia en muestras finitas. Ellas son llamadas componentes principales a un lado(cara)(ODCP), porque solo usan retardos de la serie para que sean diferenciadas de las definidas anteriormente. Veamos como son. Sean c_1 y c_2 dos enteros no negativos. Ahora sea el vector $\underline{a}' = (\underline{a}'_0, \dots, \underline{a}'_{c_1})$ con $\underline{a}'_h = (a'_{h,1}, \dots, a'_{h,k})$. Sea también B una matriz tal que $B' = [\underline{b}_0, \dots, \underline{b}_{c_2}]$, donde cada $\underline{b}_i \in \mathbb{R}^k$. Se define la primera componente principal dinámica, como

$$f(\underline{\hat{a}}) = \sum_{h=0}^{c_1} Z'_{t-h} \underline{\hat{a}}_h, \quad t = c_1 + 1, \dots, T, \quad (13.13)$$

el cual es una combinación lineal de valores retardados de las series, que tiene la propiedad de la reconstrucción de los datos originales vía

$$Z_t^{R,1}(\underline{a}, B) = \sum_{h=0}^{c_2} \underline{b}_h f_{t-h}(\underline{a}), \quad t = c_1 + c_2 + 1, \dots, T, \quad (13.14)$$

la cual minimiza el MSE en la reconstrucción. Las estimaciones $\hat{\underline{a}}$ y \hat{B} son tales que

$$(\hat{\underline{a}}, \hat{B}) \arg \min_{\|\underline{a}\|=1, B} \frac{1}{T^* k} \sum_{t=(c_1+c_2)+1}^T \|Z_t - Z_t^{R,1}\|^2, \quad (13.15)$$

con $T^* = T - (c_1 + c_2)$. Note que toda(e.d para todo t) la reconstrucción no puede llevarse a cabo.

La segunda ODCP se define como una combinación lineal de la serie retardada que puede ser usada para reconstruir de forma óptima los residuales de la primera componente. Las componentes de orden superior son definidas de forma análoga. Finalmente, la reconstrucción de Z_t usando las primeras v ODPC como

$$\underline{Z}_t^v = \sum_{i=1}^v \sum_{h=0}^{c_2^i} \hat{b}_h^i \hat{f}_{t-h}^i, \quad c_{max}^i + 1 \leq t \leq T$$

con c_1^i, c_2^i son los valores c_1 y c_2 para la i -ésima ODCP y $c_{max}^i = \max\{c_1^i + c_2^i | i \in \{1, \dots, v\}\}$.

El algoritmo secuencial para computar las ODCP se describe en [30] página 319.

13.5.2 Selección del Modelo y Pronóstico

El número de retardos, en cada componente requiere ser seleccionado. Por lo pronto, y por simplicidad, vamos a asumir que $c_1 = c_2$, esa decir, el número de retardos \underline{Z}_t usados para definir la i -ésima componente principal, y el número de retardos de \hat{f}_t^i usados para reconstruir la serie original es el mismo. Una forma de hacerlo consiste en elegir el número de componentes y el número de retardos consiste en minimizar el error de pronóstico vía validación cruzada. Supongamos que queremos hacer pronóstico h -pasos adelante. Fije un número de retardos G_{max} y un tamaño de ventana w . Entonces, dado un $g \in \{1, \dots, G_{max}\}$, computemos la primera ODCP con g retardos, usando los periodos $1, \dots, T - h - t + 1$ para $t = 1, \dots, w$, y para cada uno de esos ajustes, compute los pronósticos h -pasos adelante, y el respectivo MSE $E_{t,h}$. Estime el error de pronóstico de la validación cruzada por medio de:

$$\hat{MSE}_{1,g} = \frac{1}{w} \sum_{t=1}^w E_{t,h}.$$

Se escoge para la primera componente principal el valor $g^{*,1}$ que minimiza $\hat{MSE}_{1,g}$. Luego, fijamos la primera componente principal computada con $g^{*,1}$ retardos y repita el procedimiento con la segunda componente. Si el error de pronóstico de la validación cruzada usando las dos componentes, $\hat{MSE}_{2,g}$, es mas grande que en el que se usa una sola componente $\hat{MSE}_{1,g}$, nos detenemos y el ODCP con una componente y $g^{*,1}$ retardos, en otro caso debemos añadir la segunda componente definida usando $g^{*,2}$ retardos y se procede como antes.

El mismo procedimiento de paso a paso se puede aplicar tal que se minimice un criterio de información en vez de el error de pronóstico de la validación cruzada. El criterio es como sigue:

$$BNG_{1,g} = \log(\hat{\sigma}_{1,g}) + (g+1) \frac{\log(\min(T^{*,1,g}), g)}{\min(T^{*,1,g}), g}.$$

Después de ajustar las q DPCs a los datos, cada uno con (c_1, c_2) retardos, se pueden construir los pronósticos como sigue. Asuma que se ha decidido un procedimiento para pronosticar cada una de DPCs de forma separada. Por ejemplo, se ajusta un modelo ARMA a las componentes en una forma automática. Sea $\hat{f}_{T+h|T}^i$ para $h > 0$, el pronóstico h -pasos adelante de f_{T+h}^i con información hasta el tiempo T . Entonces se puede obtener el pronóstico h -pasos adelante de \hat{Z}_T como

$$\hat{Z}_{T+v|T} = \sum_{i=1}^q \sum_{h=0}^{c_2} \hat{b}_h^i \hat{f}_{T+v-h|T}^i.$$

14

Agrupamiento o Clustering de Series de Tiempo

La idea de esta sección es dividir un conjunto de series de tiempo en grupos homogéneos con similares propiedades y también de como clasificar una serie de tiempo dentro de un cluster entre varios que estén disponibles. En el caso inicial construiremos los grupos inicialmente, es decir, consideraremos este un problema de aprendizaje no supervisado.

Para la clasificación de series de tiempo inicialmente supondremos que tenemos un conjunto de k series de tiempo (z_{i1}, \dots, z_{iT}) para $k = 1, \dots, k$ y se desea dividir estas k series de tiempo dentro de grupos o clusteres tal que :

- (i) todas las series son clasificables,
- (ii) cada serie pertenece a 1 solo un grupo,
- (iii) cada grupo es internamente lo más homogéneo posible.

Usualmente, los procedimiento para crear grupos o clustering o:

- (1) se selecciona una medida de proximidad(distancia o disimilaridad) entre dos series de tiempo y se usa esas proximidades para formar grupos o
- (2) se define un conjunto de características o variables $\underline{x}_i \in \mathbb{R}^p$, para la i -ésima serie que resume sus propiedades y usa esos p vectores de variables para llevar a cabo el agrupamiento.

Después de eso, los clusteres son construidos mediante:

- ✓ aglomeración de los datos usando ciertas proximidades entre las series como métodos jerárquicos;
- ✓ particionando los datos mediante la similaridad en grupos como en k -means;
- ✓ estimando la mixtura de varios modelos generadores de datos, incluyendo las probabilidades de la mezcla;
- ✓ proyectando los datos en un espacio mas pequeño para encontrar los clusteres.

14.1 Distancias y Disimilaridades

La idea inicial consistirá en usar una distancia o disimilaridades entre series de tiempo para luego formar una matriz de proximidades, la cual permitirá caracterizar el conjunto de series de tiempo a la mano.

14.1.1 Distancias entre Series de Tiempo

Una distancia entre dos series de tiempo \underline{x}_t y \underline{y}_t es una función $d(\underline{x}_t, \underline{y}_t)$ que cumple tres propiedades:

Propiedad 1. $d(\underline{x}_t, \underline{y}_t) \geq 0$, y toma el valor de 0 si y sólo si $\underline{x}_t = \underline{y}_t$. (no-negativa)

Propiedad 2. $d(\underline{x}_t, \underline{y}_t) = d(\underline{y}_t, \underline{x}_t)$. (simétrica)

Propiedad 3. $d(\underline{x}_t, \underline{y}_t) \leq d(\underline{x}_t, \underline{z}_t) + d(\underline{z}_t, \underline{y}_t)$. (desigualdad triangular)

Nota 14.1. Las series $\underline{x}_t = (x_1, \dots, x_T)$ y $\underline{y}_t = (y_1, \dots, y_T)$ deben estar escaladas.

Ejemplo 14.2. Un ejemplo de medida es la Euclidiana:

$$d_E(\underline{x}_t, \underline{y}_t) = \sqrt{(\underline{x}_t - \underline{y}_t)'(\underline{x}_t - \underline{y}_t)} = \sqrt{2T(1 - \hat{\rho}_{xy}(0))}.$$

Otra forma consiste para definir distancias consiste en resumir una serie de tiempo \underline{x}_t en un vector de características $\underline{\theta}_x = (\theta_{x1}, \dots, \theta_{x1}\theta_{xp})$ y definir una distancia entre los vectores resultantes. Una familia de medidas es la de Minkowski, y se puede aplicar a estos vectores de características:

$$d_m(\underline{\theta}_x, \underline{\theta}_y) = \left(\sum_{i=1}^p |\theta_{xi} - \theta_{yi}|^m \right)^{1/m},$$

para $m > 0$.

Otro forma de construir medidas entre series de tiempo consiste en comparar sus autocorrelaciones estimadas hasta el rezago h . También se puede usar las autocorrelaciones parciales o las autocorrelaciones inversas o usar valores del periodograma. Sea $\hat{\rho}_x = (\hat{\rho}_x(0), \dots, \hat{\rho}_x(h))'$ el vector de autocorrelaciones muestrales hasta le rezago h . Se define la distancia entre las dos series de tiempo como:

$$d(\underline{x}_t, \underline{y}_t; h) = \sqrt{\sum_{j=1}^h [\hat{\rho}_x(j) - \hat{\rho}_y(j)]^2}.$$

Una variante a esta medida consiste en una medida general que incluya la dependencia de las autocorrelaciones muestrales, análogo a los que sucede con la distancia de Mahalanobis:

$$d_W(\underline{x}_t, \underline{y}_t; h) = \sqrt{(\hat{\rho}_x - \hat{\rho}_y)' W (\hat{\rho}_x - \hat{\rho}_y)}.$$

Otro enfoque consiste en resumir una serie de tiempo por medio del ajuste de un modelo $AR(h)$ y usar el vector de parámetros estimados $\hat{\phi}_x = [\hat{\phi}_x(1), \dots, \hat{\phi}_x(h)]'$. La medida con base en esta consiste en

$$d_\phi(\underline{x}_t, \underline{y}_t, h) = \sqrt{\sum_{j=1}^h [\hat{\phi}_x(j) - \hat{\phi}_y(j)]^2}.$$

Otra forma de usar los parámetros, pero ahora de un modelo ARIMA en una representación AR en la definición de la medida es como sigue:

$$d_\phi(\underline{x}_t, \underline{y}_t; h) = \sqrt{(\hat{\phi}_x - \hat{\phi}_y)' V^{-1} (\hat{\phi}_x - \hat{\phi}_y)}$$

con V^{-1} siendo una matriz de la medida de la precisión promedio de los factores π .

Otro enfoque consiste en usar el periodograma normalizado $NI(\omega_j)$ para las $[T/2]$ frecuencias ω_j , y así definir la medida entre las dos series de tiempo mediante

$$d_{NP}((\underline{x}_t, \underline{y}_t) = \sqrt{\sum_{j=1}^{[T/2]} [\log NI_x(2\pi j/T) - \log NI_y(2\pi j/T)]^2}.$$

Ver ta, también Warping distance para series de tiempo. Ver Para series no estacionarias(tendencias y estacionalidad) ver explicaciones en el libro [30] Página 214.

14.1.2 Disimilaridades entre Series de Tiempo Univariadas

Un problema con las medidas es que ellas depende de la escala de las variables y no son invariantes ante transformaciones monótonas. Por ejemplo, la distancia Euclíadiana al cuadrado no satisface la desigualdad triangular. Por eso medidas de proximidad basadas en el ordenamiento de las observaciones , no se ven afectadas por transformaciones monótonas, y se usan con frecuencia en el agrupamiento.

Una función de dos series de tiempo es una disimilaridad si verifica que

1. $D(\underline{x}_t, \underline{y}_t) \geq 0$ y $D(\underline{x}_t, \underline{y}_t) = 0$ si solo si $\underline{x}_t = a\underline{y}_t + b$,
2. $D(\underline{x}_t, \underline{y}_t) = D(\underline{y}_t, \underline{x}_t)$.

La función $D(,)$ no debe satisfacer la desigualdad triangular. Dada una medida de disimilaridad, se puede construir una medida de similaridad como sigue: Sea M la máxima disimilaridad entre dos series de tiempo en un conjunto de k series de tiempo, la función

$$S(\underline{x}_t, \underline{y}_t) = M - D(\underline{x}_t, \underline{y}_t)$$

es siempre positiva y simétrica, y se le llama medida de similaridad. Usualmente $0 \leq S(\underline{x}_t, \underline{y}_t) \leq 1$, con lo cual si toma el valor de cero quiere decir que son idénticas, mientras que si toma el valor de 1 son independientes, así las similaridades también están entre 0 y 1.

Una medida de disimilaridad basada en las correlación total($TC_{x,h} = 1 - |R_h|^{1/(h+1)}$) de una serie de tiempo estacionaria fue propuesta por Peña y Rodríguez en 2002. Esta se define como:

con las

$$D_U(\underline{x}_t, \underline{y}_t, h) = |TC_{x,h} - TC_{y,h}| = \left| |R_{x,h}|^{1/(h+1)} - |R_{y,h}|^{1/(h+1)} \right|.$$

En forma general, la matriz de proximidad D se construye como sigue. Sea $d(i, j)$ la medida elegida para definir la proximidad entre las series i y j . Luego construya una matriz cuadrada y simétrica con ceros en la diagonal, y fuera de la diagonal, en la entrada i, j igual a $d(i, j)$.

Ejemplo 14.3. Vamos a usar los datos que contienen las mediciones cada hora del $PM_{2.5}$ de los dispositivos Air-Box para marzo de 2017. Un total de 744 observaciones en el tiempo y en 516 diferentes localizaciones en Taiwan.

14.2 Agrupamiento o Aglomeración Jerárquico

Los métodos de agrupamiento jerárquico empiezan desde la matriz de proximidades(distancias o disimilaridades), D , entre las series, y con base en esta matriz y una jerarquía se construyen los grupos. Hay dos tipos de algoritmos, los aglomerativos y divisivos.

Aglomerativos. Empieza con muchos grupos, tanto como número de series hayan. En cada paso los grupos se unen usando proximidad.

Divisivos. Empieza con todas las series en un solo grupo, y en paso se hacen divisiones sucesivas hasta que las series individuales son alcanzadas.

Los dos métodos usan algún criterio para obtener el número de grupos. Para muchas series de tiempo, los métodos aglomerativos requieren de menor cómputo.

Un método aglomerativo jerárquico sigue tres pasos:

Paso 1.) Defina como elementos a ser agrupados las k series de tiempo originales, y construya la matriz de proximidad D , entre los elementos u y v , donde $u, v = 1, \dots, k$. Establezca $i = 1$ y $D_i = D$. Denotemos el (u, v) -elemento de D_i como $d_{i(uv)}$.

Paso 2.) Encuentre la distancia mínima en D_i , digamos $d^{(i)} = \min_{u,v} \{d_{i(uv)}\}$. Escoja un par de elementos tal que su distancia es igual al mínimo $d^{(i)}$. Agrupe el par para formar un nuevo elemento(nuevo grupo). Compute las distancias entre el nuevo elemento y los otros elementos. El número total de elementos es reducido por 1. Avance i en 1, y denote la matriz de distancia actualizada también por D_i .

Paso 3.) Vuelva al paso 2 con la matriz actualizada D_i y repita el procedimiento de unir hasta que el número de elementos quede en 1.

Nota 14.4. *Este procedimiento toma k iteraciones para completarse y la sucesión de distancias mínimas son denotadas por $d^1 \leq d^{(2)} \leq \dots \leq d^{(k)}$. Use esas distancias después para identificar el número de clusteres después.*

14.2.1 Criterio para Definir Distancias entre Grupos

Supongamos que tenemos un elemento A , el cual es un grupo con $n_a \geq 1$ series de tiempo, y otro grupo B con $n_b \geq 1$ series de tiempo. Ellos se unen para formar un nuevo elemento, (AB) , el cual tiene $n_{ab} = n_a + n_b$ series de

tiempo. La distancia entre este nuevo elemento, (AB) y otro elemento C , el cual tiene n_c series, se calcula siguiendo tres métodos:

El vecino mas cercano: La distancia entre C y (AB) se define como

$$d(C; AB) = \min\{d_{CA}, d_{CB}\}.$$

El vecino mas lejano: La distancia entre C y (AB) se define como

$$d(C; AB) = \max\{d_{CA}, d_{CB}\}.$$

Medición de inculación promedio: La distancia entre C y (AB) se define como

$$d(C; AB) = \frac{\sum_{c \in C} \sum_{g \in AB} d_{cg}}{n_c n_{sb}}.$$

14.2.2 Selección del número de Grupos

- ✓ Gráfico de altura y paso
- ✓ El estadístico Silhouette
- ✓ El estadístico Gap

Bibliografía

- [1] H. Akaike. *Information theory and an extension of maximum likelihood principle*. Springer, first edition, 1973.
- [2] P.J. Brockwell and Davis. R.A. *Time Series: Theory and Methods*. Springer, second edition, 2006.
- [3] P.J. Brockwell and Davis. R.A. *Introduction to Time Series and Forecasting*. Springer, third edition, 2016.
- [4] J. Brown, R.L. Durbin and J.M Evans. Techniques for testing the constancy of regression relationships over time. *Journal of the Royal Statistical Society. Series B (Methodological)*, 37(2):149–192, 1975.
- [5] J.G. De Gooijer. *Elements of Nonlinear Time Series Analysis and Forecasting*. Springer, first edition, 2017.
- [6] I. Dixon, M.F. Halperin and P. Bilokon. *Machine Learning in Finance From Theory to Practice*. Springer, first edition, 2020.
- [7] G. Dong and H. Liu. *FEATURE ENGINEERING FOR MACHINE LEARNING AND DATA ANALYTICS*. CRC Press, first edition, 2018.
- [8] E. Douc, R. Moulines and D.S. Stoffer. *Nonlinear Time Series: Theory, Methods, and Applications with R Examples*. CRC Press, first edition, 2014.
- [9] J. Durbin and D.S. Koopman. *Time Series Analysis by State Space Methods*. Oxford University Press, second edition, 2012.

- [10] D. Edgerton and C. Wells. Critical values for the cusumsq statistic in medium and large sized samples. *Oxford Bulletin of Economics and Statistics*, 56(3):355–365, 1994.
- [11] W.A. Fuller. *Introduction to Statistical Time Series*. Wiley Series in Probability and Statistics, second edition, 1996.
- [12] T. Gneiting. Making and evaluating point forecasts. *Journal of the American Statistical Association*, 106(494):746–762, 2011.
- [13] Y. Goodfellow, I. Bengio and A. Courville. *Deep Learning*. MIT Press, Cambridge, Massachusetts, 2016.
- [14] R. Hastie, T. Tibshirani and J. Friedman. *The Elements of Statistical Learning Data Mining, Inference, and Prediction*. Springer, second edition, 2009.
- [15] Y. Hmamouche, P. Przymus, A. Casali, and L. Lakhal. Gfsm: a feature selection method for improving time series forecasting. *International Journal on Advances in Systems and Measurements*, 10(3-4):255–264, 2017.
- [16] R.A. Horn and C.R. Jhonson. *Matrix Analysis*. Cambridge University Press, second edition, 2013.
- [17] M. Hornik, M. Stinchcombe and H. White. Multilayer feedforward networks are universal approximators. *Neural Networks*, 2(5):359–366, 1989.
- [18] R. Bitelli M. Huffaker and R. Rosa. *Nonlinear Time Series with R*. Oxford University Press, first edition, 2017.
- [19] R.J. Hyndman and G. Athanasopoulos. *Forecasting: principles and practice*. OTexts: Melbourne, Australia, 2021.
- [20] D. James, G. Witten, T. Hastie, and R. Tibshirani. *An Introduction to Statistical Learning with Application in R*. Springer, first edition, 2013.
- [21] S. Janson. *Gaussian Hilbert Spaces*. Cambridge University Press, first edition, 1997.
- [22] Michael Kohler and Adam Krzyzak. On the rate of convergence of a deep recurrent neural network estimate in a regression problem with dependent data, 2020.

- [23] A. Kratsios. The universal approximation property. *Annals of Mathematics and Artificial Intelligence*, 89:435–469, 2021.
- [24] Rami. Krispin. *Hands-On Time Series Analysis with R*. Packt Publishing, first edition, 2019.
- [25] M. Kuhn and K. Johnson. *Feature Engineering and Selection A Practical Approach for Predictive Models*. CRC Press, first edition, 2020.
- [26] Francesca. Lazzeri. *Machine Learning for Time Series Forecasting with Python*. John Wiley and Sons, Inc., first edition, 2020.
- [27] H. Lütkepohl. *New Introduction to Multiple Time Series Analysis*. Springer, first edition, 2005.
- [28] Wes. McKinney. *Python for Data Analysis*. O'Reilly Media, second edition, 2018.
- [29] D. Peña. *Análisis de Series de Temporales*. Alianza Editorial, segunda edición, 2010.
- [30] D. Peña and R. Tsay. *Statistical Learning for Big Dependent Data*. Jhon Wiley and Sons, Inc., first edition, 2021.
- [31] D.S.G. Pollock. *A Handbook of Time-Series Analysis, Signal Processing and Dynamics*. Academic Press, first edition, 1999.
- [32] S. Ritcher. *Statistisches und maschinelles Lernen*. Springer Spektrum, first edition, 2019.
- [33] S. Ritcher. *Statistical Analysis of Machine Learning Algorithms*. Universität Heidelberg, first edition, 2021.
- [34] Robert H. Shumway and David S. Stoffer. *Time Series Analysis and Its Applications With R Examples*. Springer, fourth edition, 2017.
- [35] Ingo Steinwart, Don Hush, and Clint Scovel. Learning from dependent observations. *Journal of Multivariate Analysis*, 100(1):175–194, 2009.
- [36] R.S. Tsay. *Analysis of Financial Time Series*. Wiley Series, third edition, 2010.
- [37] R.S. Tsay. *Multivariate Time Series Analysis with R and Financial Applications*. Wiley Series, first edition, 2014.

- [38] R.S. Tsay and R. Chen. *Nonlinear Time Series Analysis*. Wiley Series, first edition, 2019.
- [39] V. N. Vapnik. *Statistical learning theory*. Wiley-Interscience, first edition, 1998.
- [40] V. N. Vapnik. *The Nature of Statistical Learning Theory*. Springer, second edition, 2000.
- [41] W.S. Wei. *Multivariate Time Series Analysis and Applications*. Wiley, first edition, 2019.
- [42] D. H. Wolpert. The lack of a priori distinction between learning algorithms. *Neural Computationa*, 8(7):1341–1390, 1996.
- [43] I. Yeo and Jhonson R.A. A new family of power transformations to improve normality or symmetry. *Biometrika*, 87(4):954–959, 2000.