Trabajo Fin de Máster

Previsión de variables de negocio y generación de escenarios macroeconómicos con modelos estadísticos

Sandra Amado Souto

Máster en Técnicas Estadísticas

Curso 2017-2018

Propuesta de Trabajo Fin de Máster

Título en galego: Previsión de variables de negocio e xeración de escenarios macroeconómicos con modelos estatísticos

Título en español: Previsión de variables de negocio y generación de escenarios macroeconómicos con modelos estadísticos

English title: Business variable forecasting and macroeconomic scenarios generation using statical models

Modalidad: Modalidad B

Autora: Sandra Amado Souto, Universidad de Santiago de Compostela

Director: Ricardo Cao Abad, Universidad de A Coruña

Tutora: Belén María Fernández de Castro, ABANCA

Breve resumen del trabajo:

Desde el área de Planificación Estratégica y PMO de ABANCA están interesados en desarrollar modelos estadísticos que permitan pronosticar la evolución de distintas variables de negocio del sistema financiero, a partir de variables macroeconómicas o de entorno. En este ámbito de estudio, surge la necesidad de un modelo para obtener las proyecciones de uno de los indicadores financieros más utilizados, el Producto Interior Bruto, el cual se publica con un desfase de 50 días tras finalizar el trimestre. En este trabajo se intentará obtener una predicción de dicho indicador a partir de datos internos de la entidad bancaria.

Don Ricardo Cao Abad, Catedrático de la Universidad de A Coruña, doña Belén María Fernández de Castro, Especialista de ABANCA, informan que el Trabajo Fin de Máster titulado

Previsión de variables de negocio y generación de escenarios macroeconómicos con modelos estadísticos

fue realizado bajo su dirección por doña Sandra Amado Souto para el Máster en Técnicas Estadísticas. Estimando que el trabajo está terminado, dan su conformidad para su presentación y defensa ante un tribunal.

En A Coruña, a 5 de Septiembre de 2018.

El director:

Don Ricardo Cao Abad

La tutora:

Doña Belén María Fernández de Castro

La autora:

Doña Sandra Amado Souto

Agradecimientos

Agradecer, en primer lugar, a ABANCA por concederme la oportunidad de tener un primer contacto con el mundo laboral y a todos mis compañeros, en especial a Teresa Veiga Rodríguez por su ayuda y apoyo, y a mi tutora en ABANCA, Belén María Fernández de Castro, por la ayuda brindada durante las prácticas y por los conocimientos que he adquirido a lo largo de este tiempo.

En el ámbito académico, agradecer a Ricardo Cao Abad por el apoyo, la ayuda brindada durante este tiempo y, sobre todo, por los conocimientos que he adquirido gracias a la realización de este TFM.

Índice general

Resumen

Resumen en español

El PIB, Producto Interior Bruto, es el valor total de los bienes y servicios producidos en un país durante un período específico, siendo este, uno de los indicadores más utilizados en la macroeconomía ya que tiene como objetivo principal medir la actividad económica. Pero este indicador se da a conocer con un retraso de unos 50 días tras finalizar el trimestre, por lo que es de gran interés obtener un modelo que proporcione proyecciones de dicho valor.

El principal objetivo de este trabajo es obtener un modelo, en el cual, a través de datos de la entidad bancaria, se pueda predecir el valor del Producto Interior Bruto, centrándonos para ello en Galicia, pues es la principal área de desarrollo de la actividad de ABANCA. A lo largo del documento se estudiarán distintas metodologías, las cuales se aplicarán a los datos que vamos a manejar para llegar así al objetivo propuesto.

Esta memoria se organizará de la siguientes manera: en la primera parte se motivará el problema a estudiar y se introducirán las variables que se utilizarán a lo largo del documento y la metodología que se empleará (componentes principales, componentes principales dinámicas y distintos modelos de regresión) y en la segunda parte se aplicará la metodología explicada en la parte I y se compararán los distintos modelos de regresión.

English abstract

The GDP, gross domestic product, is the total values of goods and services produced in a country during a specific period. It is the most important macroeconomic indicator and its main aim is to measure economic activity. However, this indicator is published with a difference of 50 days after the end of the quarter. Therefore, it is of great interest to obtain a model, which provides projections of futures values of GDP.

The main objective of this academic work is to obtain a model, in which, through data from a bank, the value of the GDP can be predicted. We will focus on Galicia, the main area of development of ABANCA. During the document, different methodologies will be studied.

This document will be organized as follows: in the first part, the problem will be motivated and the variables and methodology will be introduced (main components, main dynamic components and different regression models). In the second part, the methodology explained in the first part will be applied to real data. Different regression models have been compared.

Parte I

Metodología

Capítulo 1

Motivación e información disponible

Este es un capítulo inicial, en el cual se va a explicar la importancia de este trabajo para ABANCA y las variables utilizadas a lo largo de la memoria. En la sección 1.1 se explicará la motivación del problema, es decir, la importancia del problema que se considerará, el objetivo que se persigue y qué tipo de información interna puede ayudar al objetivo. En la sección 1.2 se definirá el Producto Interior Bruto, conocido como PIB, siguiendo para ello, Mochón y Beker (2006) [8] y en la sección 1.3 se introducirán las variables internas de ABANCA.

1.1. Motivación

Uno de los indicadores más importantes de la economía es el Producto Interior Bruto, conocido como PIB, el cual nos indica cómo evoluciona la economía en un territorio y tiene una frecuencia trimestral. Sin embargo, este valor no se publica al terminar el trimestre al que hace referencia, sino, que se publica con un desfase, más o menos, de 50 días. Por ejemplo, el PIB del tercer trimestre del 2018 se publicará el 29 de Noviembre de dicho año. Es por este motivo que surge la necesidad de un modelo para poder obtener proyecciones de dicho indicador.

Muchos de los modelos empleados para obtener estas proyecciones utilizan variables externas. Sin embargo, en este trabajo lo que se pretende es obtener proyecciones del PIB utilizando para ello variables internas de la propia entidad.

Dado que ABANCA es la entidad de referencia para el 65 % de la población gallega y que las familias gallegas tienen en la entidad el 42 % de sus depósitos y el 33 % de sus créditos, se puede asegurar que los datos internos de la entidad son una muestra representativa de lo que ocurre en la comunidad autónoma. Lo que hace pensar que estos datos pueden ser una fuente de información fiable como termómetro de la economía gallega.

Concretamente, las variables internas utilizadas sólo harán referencia al sector familias. Por lo que, inicialmente, se estudiará la relación entre la evolución de las variables internas y la componente de consumo de los hogares del PIB. Pero dado que esta componente tiene un peso muy relevante en la evolución del PIB, históricamente el 60 % de la variación del PIB se debe a la variación de la componente de consumo de los hogares, se ha abordado también el estudio de la relación entre las variables internas de la entidad y el PIB de Galicia.

Para llevar a cabo el citado análisis se considerarán trece variables. Por ese motivo se reducirá la dimensión de dicho conjunto utilizando para ello el análisis de componentes principales dinámicas, pues nuestro conjunto de variables explicativas son series temporales. Una vez reducida la dimensión se ajustará un modelo de regresión, donde la variable respuesta será el PIB o el consumo en hogares gallegos y las variables explicativas las componentes principales dinámicas obtenidas anteriormente. Una vez ajustado

el modelo se deberá chequearlo para comprobar si este es válido y poder realizar predicciones a futuro.

Finalmente se intentará extrapolar el modelo utilizado para la previsión del PIB y del consumo de hogares gallegos a cada provincia, para poder observar la evolución de dichas variables. Está será una primera aproximación, quedando para el futuro poder mejorar dicho modelo.

1.2. PIB: variable respuesta

En la macroeconomía es necesario un valor que permita obtener una visión global de la economía. La medición de la actividad económica a lo largo de un período solo ha sido posible gracias a la evolución de la contabilidad nacional. De los distintos agregados que recoge la contabilidad nacional, el más significativo es el Producto Interior Bruto (PIB).

Definición 1.2.1. *El **Producto Interior Bruto** (PIB) es una cantidad que mide el valor monetario total de los bienes y servicios finales producidos para el mercado, dentro de las fronteras de un país, en un año dado.*

Se analizará, siguiendo Mochón y Beker (2006) [8], cada parte de la definición del PIB.

El valor monetario total ...

Toda la economía produce miles de bienes y servicios distintos y cada uno se mide en una unidad diferente. Para poder combinarlos en una única cifra, se sumará el número de unidades monetarias (euros) por el cuál se vende cada bien o servicio.

... de todos los bienes y servicios finales ...

Cuando se mide el PIB, no se cuentan todos los bienes y servicios producidos en el país, sino únicamente los que se venden a los usuarios finales, pues si se incluyeran bienes intermedios se estarían contando dos veces, puesto que en el precio final ya van incluidos los precios de las demás etapas. Para entenderlo mejor se puede consultar el ejemplo del cuadro 13.1 de Mochón y Beker (2006, cap. 13) [8].

... producidos ...

En la medición del PIB solo se incluyen los bienes y servicios producidos en el período considerado. Las compras a terceros o actividades financieras (acciones o bonos) no se incluyen en el PIB. Pues:

- Bonos y acciones: representan un derecho de propiedad o a recibir pagos en el futuro, pero no son bienes y servicios.
- Compra a terceros: bienes que fueron producidos, pero no en el periódico que se está a considerar.

... para el mercado ...

El PIB no incluye todos los bienes y servicios producidos en la economía, sino solo los que se producen para el mercado, esto es, con la intención de ser vendidos.

... durante un año dado ...

El PIB es una variable de flujo que mide un proceso que se desarrolla a lo largo de un período.

... dentro de las fronteras del país

El PIB mide la producción dentro de las fronteras de un país, independientemente de que haya sido producida, o no por ciudadanos de dicho país. Esto significa que se incluyen los productos fabricados con recursos que son propiedad de extranjeros y por personas extranjeras que residen en dicho país.

Para entender mejor esto último se utilizará el ejemplo de Mochón y Beker (2006, cap. 13) [8]. Si un equipo de fútbol de un país se va ha hacer una gira por España, el valor de los servivios, que en este caso son extranjeros, se incluye en el PIB de España y si una persona extranjera está trabajando en Estaña va a formar parte del PIB español y no del PIB de su país de origen.

1.2.1. Cálculo del PIB

Para el cálculo del PIB se verán tres métodos.

1. **Cálculo del PIB por el método del gasto**. Para poder realizar este cálculo, se divide la producción en cuatro categorías según el grupo de la economía que la compre.

   - **Consumo de bienes y servicios (C)**: comprados por las familias. El consumo es el gasto de bienes y servicios realizado por las familias. Sin embargo, no incluye el flujo de servicios prestados a lo largo de su vida útil. Esto quiere decir, el consumo incluye todo lo que las familias compren en el período considerado durante un año (alimentos, ropa, ...). Sin embargo, hay dos clases de bienes que compran las familias en dicho período y no forman parte del consumo puesto que no son producidas en el período a considerar. Estos serían los bienes usados (vehículos de segunda mano o libros usados) y los activos (acciones, bonos, ...).

     Este es el elemento más importante del PIB, pues representa un alto porcentaje del valor del PIB.
   - **Inversión privada en bienes y servicios (I)**: compradas por las empresas. Se obtiene sumando:
     - Compras empresariales de planta y equipo.
     - Construcción residencial.
     - Variación de existencias. Esto es, los bienes que se han producido pero que aún no han sido vendidos. Son considerados puesto que, forman parte del stock de capital del país y ofrecerán servicios en el futuro, una vez que se vendan y se utilicen.
   - **Gasto público (G)**: bienes y servicios comprados por el sector público. El gasto público es el realizado por el sector público en bienes y servicios, y comprende todos los gastos en que este sector incurre para pagar la nómina de sus empleados más los costos de los bienes (carreteras, ferrocarriles, ...) y servicios (de consultoría, financieros,...) que compra el sector privado.
   - **Exportaciones netas (XN)**: bienes y servicios comprados por los extranjeros, menos las importaciones, es decir, $exportaciones - importaciones$.

   Por lo tanto:

   $$\mathbf{PIB=C+I+G+XN}$$

2. **Cálculo del PIB por el método del valor agregado**. El valor del PIB se obtiene sumando el valor agregado, diferencia entre el precio de venta de un bien y el costo de los bienes intermedios necesarios para producirlos, que generan distintas ramas de la economía:

   - Agricultura, ganadería, silvicultura y pesca.
   - Industria.
   - Construcción.
   - Servicios.

   Para entender mejor este método, se ilustrará con el ejemplo del cuadro 13.2 de Mochón y Beker (2006, cap. 13) [8], en el cual se considera la producción de 30 piezas de pan, desde que el agricultor produce el trigo hasta que se vende en la panadería, cuadro 1.1.

Empresas (producción)	Costo factores (productos intermedios)	Precio venta	Valor agregado
Agrícola	0	5	5-0=5
Harinera	5	15	15-5=10
Panadera	15	25	25-15=10
Distribuidora	25	36	36-25=11
			Total=5+10+10+11=36

Cuadro 1.1: Cálculo por el método del valor agregado

3. **El PIB por el método de los costos**. El PIB, por este método, se obtiene sumando los ingresos (ingresos del trabajo, alquileres,...) que han obtenido todas las familias. Es decir:

PIB= sueldos, salarios y otros ingresos del trabajo +
intereses, alquileres y otros ingresos de la propiedad +
impuestos indirectos +
depreciación o amortización +
beneficios

1.2.2. Obtención de la serie del PIB y del consumo de hogares gallegos

La serie del PIB gallego se obtiene de la página web del Instituto Gallego de Estadística (IGE), el cual lo calcula por el método del gasto y su frecuencia es trimestral. Se va a trabajar con la serie del PIB real, en volumen, con base 2010 corregida por calendario y estacionalidad. Además, para poder comparar con el PIB publicado se tiene que calcular la variación interanual del PIB trimestral. Para ello se comienza calculando la variación trimestral del PIB de la siguiente forma:

$$\left(\frac{Q_i^t}{Q_i^{t-1}} - 1\right) \times 100$$

donde Q_i^t con $i = 1, 2, 3, 4$ es el PIB del trimestre i del año t. De modo que si quisiéramos obtener la variación del primer trimestre del 2018, se tendría que realizar el cálculo:

$$\left(\frac{Q_1^{2018}}{Q_1^{2017}} - 1\right) \times 100$$

La variación iteranual del PIB se obtiene realizando un promedio de las variaciones de los cuatros trimestres del año a considerar.

De forma análoga, se obtiene la serie del consumo de hogares gallegos.

1.3. Variables explicativas

En esta sección se presentarán las variables explicativas que se utilizarán a lo largo de la memoria, pero antes de eso, se hará una pequeña introducción al lenguaje SQL, pues el sistema que administra la base de datos interna de ABANCA, Data Ware House (DWH), emplea dicho lenguaje para explorar los datos almacenados en ella. La información se almacena en forma de tabla y cada tabla tiene un nombre único en dicha base de datos.

La estructura de una consulta SQL es:

select *Campos separados por comas, es decir, las columnas que se quieren seleccionar o ∗ para seleccionar todos los campos*
from *Tabla*
where *Condición 1* **and** *Condición 2* **and** ...

ID-EMPLEADO	NOMBRE	APELLIDOS	SEXO	CARGO	ASIGNATURA	SALARIO
1	Carlos	Rubio Martínez	H	Profesor	Inglés	1658
2	Antía	Agro Castro	M	Profesora	Matemáticas	1300
3	Paula	López Camaño	M	Profesora	Lengua castellana	1800
4	Jorge	Antelo Pérez	H	Profesor	Francés	1200

Cuadro 1.2: Tabla EMPLEADOS

Supongamos que tenemos una tabla, la cual es ficticia, cuyo nombre es EMPLEADOS, cuadro 1.2, y se quiere consultar qué empleados tienen un salario mayor a 1600€. Entonces la consulta SQL es:

select *nombre, apellidos*
from *empleados*
where *salario*> 1600

Nótese que al realizar la consulta se ha escrito en minúsculas cuando el nombre de las columnas y de la tabla están en mayúscula. Esto se puede hacer puesto que no las diferencia. La salida de esta consulta se muestra en el cuadro 1.3.

NOMBRE	APELLIDOS
Carlos	Rubio Martínez
Paula	López Camaño

Cuadro 1.3: Consulta

Supongamos que se tienen tres tablas y se quiere unir esa información (se pueden unir más). Para poder unirlas, las tres tablas tendrán que tener, al menos, una columna que contenga la misma información. La estructura de la consulta SQL sería:

select *Campos separados por comas, es decir, las columnas que se quieren seleccionar o ∗ para seleccionar todos los campos*
from *Tabla 1* **inner join** *Tabla 2*
on *Nombre de la columna por las que se van a unir Tabla 1 y Tabla 2*
inner join *Tabla 3*
on *Nombre de la columna por las que se van a unir Tabla 1 o 2 y Tabla 3*
where *Condición 1* **and** *Condición 2* **and** ...

NOMBRE	APELLIDOS	ASIGNATURA	ID-PROFESOR
María	López Vega	Matemáticas	2
Juán	Mártinez Alonso	Francés	4
Luís	Ferreiro García	Lengua castellana	3
Marcos	García García	Inglés	1

Cuadro 1.4: Tabla ALUMNOS

Para entender mejor lo que hace *inner join... on* se realizará un pequeño ejemplo donde los datos son ficticios. Para ello se considerará el cuadro 1.4, en la cual, un alumno está matriculado en una asignatura y a cada asignatura se le ha asociado un profesor. Nótese que no conocemos el nombre del profesor si no la identificación de cada uno de ellos y lo que se quiere saber es qué profesor será asignado a cada alumno. Para ello se deberá unir las tablas de los cuadros 1.2 y 1.4 a través de la columna ID PROFESOR. De modo, que la consulta SQL sería:

```
select a.nombre as alumno , a.apellido as apellido-alumno, a.asignatura, e.nombre as profesor,
e.apellido as apellido-profesor
from empleados e inner join alumnos a
on a.id-profesor=e.id-empleado
```

En la consulta se ha escrito *empleados e* y *alumnos a*, es decir, a partir de ahora la tabla empleados se identificará con la abreviatura e y la tabla alumnos con la abreviatura a. Para evitar la ambigüedad, debido a que ambas tienen columnas con el mismo nombre, es necesario indicar en *select* de qué tabla procede cada columna que queremos que nos devuelva la consulta y con *as* se le asigna un nuevo nombre a la columna de la tabla que nos devuelve la consulta. La salida se muestra en el cuadro 1.5.

alumno	apellido-alumno	ASIGNATURA	profesor	apellido-profesor
María	López Vega	Matemáticas	Antía	Agro Castro
Juán	Mártinez Alonso	Francés	Jorge	Antelo Pérez
Luís	Ferreiro García	Lengua castellana	Paula	López Camaño
Marcos	García García	Inglés	Carlos	Rubio Martínez

Cuadro 1.5: Consulta

En todos los ejemplos hemos seleccionado columnas con el comando **select** pero también se pueden incluir operaciones en esta línea. Por ejemplo, supongamos que lo que queremos saber es cuánto cobran todos los profesores juntos. Entonces, la consulta a realizar es:

```
select sum(e.salario) as sum-salario
from empleados e
```

La cual nos devolvería el cuadro 1.6.

sum-salario
5958

Cuadro 1.6: Consulta

También podemos saber la suma de los salarios según el sexo del profesor. Para ello, se utilizará el comando **group by** y la consulta que habría que realizar es:

```
select e.sexo as sexo, sum(e.salario) as sum-salario
from empleados e
group by e.sexo
```

Siendo la salida el cuadro 1.7.

sexo	sum-salario
H	2858
M	3100

Cuadro 1.7: Consulta

Alguna de las operaciones que se pueden realizar en la línea **select** son:

- Operaciones: +, -, ×, /.
- sum: para sumar.
- count: para contar.
- extract: extraer una parte de una variable. Supongamos que se tiene una columna con el nombre **FECHA** y cuya estructura es *DAY-MONTH-YEAR*. Para extraer el año de la fecha basta con escribir: *extract(year from fecha)*.

Realizando varias consultas en DWH y utilizando IBM SPSS Modeler para fusionar, seleccionar parámetros y ordenar las tablas, se han obtenido las variables explicativas que se emplearán. Dichas variables, referidas a Galicia y con frecuencia mensual, comprendidas entre Julio del 2007 y Marzo de 2018, son:

1. **Gasto con tarjetas de débito**: : importe (€) acumulado en el mes de los pagos realizados con tarjeta de débito por los clientes del segmento familias de la entidad.
2. **Saldo objeto plazo**: saldo (€) que los clientes del segmento familias de la entidad tienen en productos de plazo fijo a cierre de mes.
3. **Saldo fondo inversión**: saldo (€) que los clientes del segmento familias de la entidad tienen en fondos de inversión a cierre de mes.
4. **Saldo fondo pensiones**: saldo (€) que los clientes del segmento familias de la entidad tienen en planes de pensiones a cierre de mes.
5. **Saldo seguros ahorro**: saldo (€) que los clientes del segmento familias de la entidad tienen en seguros de inversión a cierre de mes.
6. **Saldo hipotecas**: saldo (€) que los clientes del segmento familias de la entidad tienen en préstamos con garantía hipotecaria a cierre de mes. Se incluye únicamente saldo en situación normal.

7. **Saldo consumo sin tarjeta de crédito**: saldo (€) que los clientes del segmento familias de la entidad tienen en préstamos al consumo, excluidas las tarjetas de crédito, a cierre de mes. Se incluye únicamente saldo en situación normal.

8. **Saldo dudoso hipotecas**: saldo en situación dudosa (€) que los clientes del segmento familias de la entidad tienen en préstamos con garantía hipotecaria a cierre de mes.

9. **Saldo dudoso consumo sin tarjeta de crédito**: saldo en situación dudosa (€) que los clientes del segmento familias de la entidad tienen en préstamos al consumo, excluidas las tarjetas de crédito, a cierre de mes.

10. **Número operaciones TPV**: número de operaciones, acumuladas en el mes, que se realizan en los TPVs de la entidad.

11. **Importe TPV**: importe (€) de operaciones, acumulado en el mes, que se realizan en los TPVs de la entidad.

12. **Gasto mensual tarjetas de crédito**: importe (€) acumulado en el mes de los pagos realizados con tarjeta de crédito por los clientes del segmento familias de la entidad.

13. **Número de operaciones con tarjeta de crédito**: número de operaciones, acumulado en el mes, realizadas con tarjeta de crédito por los clientes del segmento familias de la entidad.

Las trece variables están relacionadas con el consumo en hogares gallegos, por ese motivo se realizará, en primer lugar, un ajuste del consumo, en el cual la variable respuesta sera la serie gasto en hogares gallegos obtenida del IGE. Como se verá más adelante, con nuestras variables se obtiene un buen ajuste, de modo que, al ser el consumo, más o menos, un 60 % del valor del PIB, dichas variables también proporcionarán un buen ajuste del PIB.

Capítulo 2

Reducción de la dimensión de un conjunto de variables

Como ya se ha mencionado en el capítulo anterior, para la realización del trabajo se utilizarán trece variables internas, por lo que será de gran interés poder reducir la dimensión. Por ese motivo comenzaremos con una breve introducción sobre el análisis de componentes principales. Al ser nuestras variables series temporales será más conveniente la utilización de componentes principales dinámicas, (puede consultarse [7]), pues son necesarias debido a la dependencia temporal que presentan las series.

Este capítulo se organizará de la siguiente manera: en la sección 2.1, siguiendo Peña (2002) [10], se explicará el análisis de componentes principales y en la sección 2.2, siguiendo Peña y Yohai (2016) [12], se hará una introducción a las componentes principales dinámicas.

2.1. Análisis de componentes principales

El análisis de componentes principales (PCA) se considera una técnica de reducción de la dimensión, ya que permite describir el comportamiento de un gran número de variables, d, a partir de un pequeño subconjunto $r < d$ de ellas, el cual recibe el nombre de componentes principales. Por lo tanto, dadas n observaciones de d variables, se analiza si es posible representar adecuadamente esta información con un número menor de variables construidas como combinación lineales de las originales. En la reducción de la dimensión se asume una pequeña pérdida de información.

Nótese que para poder obtener las componentes principales de un conjunto debemos conocer su matriz de covarianzas o de correlación. Si no hay correlación entre las variables las componentes principales son las variables originales. Por lo que antes de la obtención de dichas componentes, tendremos que calcular la matriz de correlaciones de las variables y aplicarle posteriormente el test de esfericidad.

2.1.1. Test de esfericidad

Sea $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ una muestra aleatoria simple de una $N_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. El test de esfericidad consiste en contrastar la hipótesis nula de que las variables son incorreladas y tienen la misma varianza, por lo tanto:

$$H_0 = \boldsymbol{\Sigma} = \sigma^2 \mathbf{I_d}$$

siendo σ^2 la varianza común desconocida. Nótese que bajo la hipótesis alternativa, la matriz de covarianzas no esta sujeta a restricciones y el vector de medias carece de restricciones, tanto en la hipótesis nula como en la hipótesis alternativa.

Aplicando el procedimiento de razones de verosimilitud, el estadístico de contraste resulta ser:

$$-2 \log \lambda(\mathbf{x}) = -2 \log \frac{\sup_{\boldsymbol{\mu}} L(\mathbf{x}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2 \mathbf{I_d})}{\sup_{\boldsymbol{\mu}, \boldsymbol{\Sigma}} L(\mathbf{x}, \boldsymbol{\mu}, \boldsymbol{\Sigma})}$$

Recordemos que la función log-verosimilitud se expresa del siguiente modo:

$$\log L(\mathbf{x}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = c - \frac{n}{2}\log|\boldsymbol{\Sigma}| - \frac{1}{2}\sum_{i=1}^{n}(\mathbf{x_i} - \overline{\mathbf{x}})'\boldsymbol{\Sigma}^{-1}(\mathbf{x_i} - \overline{\mathbf{x}}) - \frac{n}{2}(\overline{\mathbf{x}} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\overline{\mathbf{x}} - \boldsymbol{\mu})$$

Además:

$$\sup_{\boldsymbol{\mu}} \log L(\mathbf{x}, \boldsymbol{\mu}, \sigma^2\mathbf{I_d}) = c - \frac{n}{2}(\log|\boldsymbol{\Sigma}| + tr(\boldsymbol{\Sigma}^{-1}\mathbf{S}))$$

y

$$\sup_{\boldsymbol{\Sigma}} \sup_{\boldsymbol{\mu}} \log L(\mathbf{x}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = c - \frac{n}{2}(\log|\mathbf{S}| + d)$$

siendo $\mathbf{S}$, en ambos casos, la estimación de la matriz de covarianzas, $\boldsymbol{\Sigma}$.

Para ver detalladamente cómo se obtiene las igualdades anteriores se puede consultar Peña (2002) [10].

En nuestro caso,

$$\sup_{\boldsymbol{\mu}} \log L(\mathbf{x}, \boldsymbol{\mu}, \sigma^2\mathbf{I_d}) = c - \frac{n}{2}(\log|\sigma^2\mathbf{I_d}| + tr((\sigma^2\mathbf{I_d})^{-1}\mathbf{S})) = c - \frac{n}{2}(\log(\sigma^{2d}|\mathbf{I_d}|) + tr(\frac{1}{\sigma^2}\mathbf{I_d}\mathbf{S})) = c - \frac{n}{2}(d\ \log(\sigma^2) + \frac{1}{\sigma^2}tr(\mathbf{S}))$$

Ahora busquemos el supremo respecto de σ^2, para lo cual, efectuaremos la derivada correspondiente:

$$\frac{\partial \sup_{\boldsymbol{\mu}} L(\mathbf{x}, \boldsymbol{\mu}, \sigma^2\mathbf{I_d})}{\partial\sigma^2} = -\frac{n}{2}(d\frac{1}{\sigma^2} - \frac{1}{\sigma^4}tr(\mathbf{S}))$$

Para calcular el máximo se iguala a cero la expresión anterior y se despeja σ^2.

$$d\frac{1}{\sigma^2} - \frac{1}{\sigma^4}tr(\mathbf{S}) = 0 \implies d\sigma^2 - tr(\mathbf{S}) = 0 \implies \sigma^2 = \frac{1}{d}tr(\mathbf{S}) = \frac{1}{d}\sum_{j=1}^{d}\lambda_j = \frac{1}{d}da_0 = a_0$$

siendo $\lambda_1, \ldots, \lambda_d$ los autovalores de $\mathbf{S}$ y a_0 su media aritmética. Entonces el estimador de máxima verosimilitud es $\widehat{\sigma}^2 = a_0$.

Después de estos cálculos, tenemos que el estadístico de contraste adopta la siguiente forma:

$$-2\log\lambda(\mathbf{x}) = -2\left(-\frac{n}{2}(d\log(a_0) + \frac{1}{a_0}tr(\mathbf{S}) - \log|\mathbf{S}| - d)\right) = n\left(d\log(a_0) + \frac{1}{a_0}tr(\mathbf{S}) - \log|\mathbf{S}| - d\right) = n\left(d\log(a_0) + \frac{1}{a_0}\sum_{j=1}^{d}\lambda_j - \log\left(\prod_{j=1}^{d}\lambda_j\right) - d\right) = n\left(d\log(a_0) + \frac{1}{a_0}da_0 - \log(g_0^d) - d\right) = n\,(d\log(a_0) + d - d\log(g_0) - d) = n\left(d\log\left(\frac{a_0}{g_0}\right)\right) = n\ d\log\left(\frac{a_0}{g_0}\right)$$

siendo g_0 la media geométrica de los autovalores de $\mathbf{S}$.

Por último, veamos qué distribución sigue nuestro estadístico. La distribución exacta de dicho estadístico bajo la hipótesis nula no está disponible. En su lugar, usaremos la distribución asintótica.

$$-2\log\lambda(\mathbf{x}) = n\ g\ \log\left(\frac{a_0}{g_0}\right) \sim \chi_m,$$

donde los grados de libertad m, pueden obtenerse tal y como se indica en el párrafo siguiente.

Recordemos que bajo la hipótesis nula el vector de medias no tenía ninguna restricción, sin embargo, la matriz de covarianzas estaba sujeta a una restricción. Por lo tanto, bajo la hipótesis nula tendremos d parámetros libres ($m_0 = d$). Por otro lado, bajo la hipótesis alternativa, tanto el vector de medias como la matriz de covarianzas no estaban sujetas a ninguna restricción. Por lo tanto, bajo la hipótesis alternativa tendremos $2d + \binom{d}{2}$ parámetros libres, $m_1 = 2d + \binom{d}{2}$. De esta forma $m = m_1 - m_0 = d + \binom{d}{2} = d + \dfrac{d!}{2!(d-2)!} = d + \dfrac{d(d-1)}{2} = \dfrac{2d + d^2 - d}{2} = \dfrac{1}{2}(d + d^2) = \dfrac{1}{2}d(d+1)$. Por lo tanto, nuestro estadístico sigue una distribución chi cuadrado con $m = \dfrac{1}{2}d(d+1)$ grados de libertad.

2.1.2. Obtención de las componentes principales

Supongamos que se dispone de d valores de n elementos de una población dispuestos en una matriz $\mathbf{X}$ de dimensión $n \times d$, donde las columnas contienen las variables y las filas los elementos de la muestra. En este apartado definiremos el concepto de componentes principales y veremos cómo se calculan a partir de la matriz de covarianzas de un vector aleatorio. Todo esto también se puede aplicar a un conjunto de datos. En este último caso, se puede pensar que el análisis se está aplicando a una distribución de probabilidades discreta equiprobable sobre los vectores observados.

Definición 2.1.1. *Sea $\boldsymbol{X} = (X_1, \cdots, X_d)'$ un vector aleatorio d-dimensional con vector de medias $\boldsymbol{\mu} = E(\boldsymbol{X})$ y matriz de covarianzas $\boldsymbol{\Sigma} = E((\boldsymbol{X} - \boldsymbol{\mu})(\boldsymbol{X} - \boldsymbol{\mu})')$. Se define la primera componente principal de $\boldsymbol{X}$ como una variable aleatoria Z_1 tal que:*

$$Z_1 = \boldsymbol{v_1'}\boldsymbol{X} = v_{11}X_1 + \cdots + v_{d1}X_d \text{ con } \boldsymbol{v_1} = (v_{11}, \ldots, v_{d1})' \in \mathbb{R}^d$$

$$Var(Z_1) = \text{máx}\{Var(\boldsymbol{v}'\boldsymbol{X}) : \boldsymbol{v} \in \mathbb{R}^n, \boldsymbol{v}'\boldsymbol{v} = 1\}$$

Así pues, la primera componente principal es una combinación lineal normalizada de las variables $\boldsymbol{X}$ y entre todas ellas es la de mayor varianza.

Teorema 2.1.1. *La primera componente principal de $\boldsymbol{X}$ adopta la forma:*

$$Z_1 = \boldsymbol{v}_1'\boldsymbol{X}$$

siendo

$$Var(Z_1) = \lambda_1$$

donde λ_1 el mayor autovalor de la matriz de covarianzas $\boldsymbol{\Sigma}$ y $\boldsymbol{v_1}$ un autovector asociado a λ_1 tal que $\boldsymbol{v_1'}\boldsymbol{v_1} = 1$, es decir, $\boldsymbol{v_1'}$ es unitario.

La demostración del teorema anterior puede consultarse en Peña (2002) [10].

Definición 2.1.2. *Se define la segunda componente principal de $\boldsymbol{X}$ como una variable aleatoria Z_2 tal que:*

$$Z_2 = \boldsymbol{v_2'}\boldsymbol{X} = v_{12}X_1 + \cdots + v_{d2}X_d \text{ con } \boldsymbol{v_2} = (v_{12}, \ldots, v_{d2})' \in \mathbb{R}^d$$

$$Var(Z_2) = \text{máx}\{Var(\boldsymbol{v}'\boldsymbol{X}) : \boldsymbol{v} \in \mathbb{R}^n, \boldsymbol{v}'\boldsymbol{v} = 1, \boldsymbol{v}'\boldsymbol{v_1} = 0\}$$

Así pues, la segunda componente principal es una combinación lineal normalizada de las variables $\boldsymbol{X}$, correspondiente a una dirección ortogonal a la de la primera componente principal, y entre todas esas combinaciones lineales formadas por vectores unitarios a $\boldsymbol{v_1}$, es el de mayor varianza.

Teorema 2.1.2. *La segunda componente principal de* $\boldsymbol{X}$ *adopta la forma:*

$$Z_2 = \boldsymbol{v}_2' \boldsymbol{X}$$

siendo

$$Var(Z_2) = \lambda_2$$

donde λ_2 *el segundo mayor autovalor de la matriz de covarianzas* $\boldsymbol{\Sigma}$ *y* $\boldsymbol{v_2}$ *un autovector asociado a* λ_2 *tal que* $\boldsymbol{v_2' v_2} = 1$ *y* $\boldsymbol{v_1' v_2} = 0$*, es decir que* $\boldsymbol{v_2'}$ *es unitario y ortogonal a* $\boldsymbol{v_1}$*.*

La demostración del teorema anterior puede consultarse en Peña (2002) [10].

De forma análoga a la anterior, se puede seguir definiendo el resto de componentes principales. Entonces, se definen las d componentes principales de $\mathbf{X}$ como las variables aleatorias $(Z_1, \cdots, Z_d)$ tales que:

$$Z_1 = \mathbf{v_1'X}, Z_2 = \mathbf{v_2'X}, \ldots, Z_d = \mathbf{v_d'X} \text{ con } \mathbf{v_1}, \mathbf{v_2}, \ldots, \mathbf{v_d} \in \mathbb{R}^d, \text{ siendo}$$

$$\begin{gathered} Var(Z_1) = \text{máx}\{Var(\mathbf{v'X}) : \mathbf{v} \in \mathbb{R}^d, \mathbf{v'v} = 1\} \\ Var(Z_2) = \text{máx}\{Var(\mathbf{v'X}) : \mathbf{v} \in \mathbb{R}^d, \mathbf{v'v} = 1, \mathbf{v'v_1} = 0\} \\ \vdots \\ Var(Z_d) = \text{máx}\{Var(\mathbf{v'X}) : \mathbf{v} \in \mathbb{R}^d, \mathbf{v'v} = 1, \mathbf{v'v_1} = 0, \mathbf{v'v_2} = 0, \ldots, \mathbf{v'v_{d-1}} = 0\} \end{gathered}$$

2.1.3. Proporción de varianza explicada por las componentes principales y criterios para reducir la dimensión

Recordemos que $Var(\mathbf{Z_1}) = \lambda_1$, por lo que la proporción de variabilidad explicada por la primera componentes principal vendrá dada por el cociente:

$$\frac{\lambda_1}{\lambda_1 + \lambda_2 + \cdots + \lambda_d}$$

De forma análoga se define la variabilidad explicada por cada una de las d componentes principales.

Si consideramos las j primeras componentes principales, la proporción de variabilidad explicada por estas viene dado por el siguiente cociente:

$$\frac{\lambda_1 + \lambda_2 + \cdots + \lambda_{j-1} + \lambda_j}{\lambda_1 + \lambda_2 + \cdots + \lambda_{j-1} + \lambda_j + \lambda_{j+1} + \cdots + \lambda_d}$$

Ahora debemos de decidir cómo vamos a reducir la dimensión, es decir, con cuántas componentes nos vamos a quedar, ya que la reducción de la dimensión conlleva una pérdida de información. Esto nos lleva a definir los siguientes criterios de decisión:

- Criterio de la varianza explicada. Consiste en seleccionar las componentes principales que expliquen conjuntamente, al menos, un 90 % o un 95 % de la variabilidad total. Esta regla es arbitraria y debe aplicarse con cierto cuidado, pues puede ocurrir que con una única componente principal se explique más del 90 % de la variabilidad pero que sea más adecuando añadir otras componentes para explicar la forma de las variables.

  Este criterio será el que utilicemos en la práctica.

- Criterio del valor propio (Kaiser). Consiste en capturar las componentes principales cuyos autovalores sean mayores que $traza(\Sigma)/d$. El motivo de tomar este umbral es que es precisamente la media de los autovalores.

- Gráfico de sedimentación. Consiste en representar los valores propios, λ_i con $i = 1, \ldots, d$, frente a i en orden decreciente y se van seleccionando las componentes hasta que las restantes tengan aproximadamente el mismo valor. la idea es buscar un "codo" en el grafo, es decir, un punto al partir del cual los valores propios son aproximadamente iguales y seleccionar aquellas que estén antes del "codo".

Para más información sobre los criterios puede consultarse Peña (2002) [10].

2.1.4. Ejemplo del cálculo de las componentes principales

En este apartado, realizaremos un análisis de componentes principales suponiendo que $\mathbf{X}$ es un vector aleatorio que sigue una distribución normal de media cero y matriz de covarianzas:

$$\mathbf{\Sigma} = \begin{pmatrix} 8 & 5 \\ 5 & 4 \end{pmatrix}$$

En primer lugar, calcularemos los autovalores de $\mathbf{\Sigma}$, es decir, tendremos que calcular $|\mathbf{\Sigma} - \lambda \mathbf{I_2}|$, igualarlo a 0 y despejar λ. Realizando las operaciones, obtenemos que los autovalores son: $\lambda_1 = 6 + \sqrt{29}$ y $\lambda_2 = 6 - \sqrt{29}$.

A partir de los autvalores calcularemos los autovectores, pues $\mathbf{\Sigma}\, \mathbf{v_1} = \lambda_1 \mathbf{v_1}$ y $\mathbf{\Sigma}\, \mathbf{v_2} = \lambda_2 \mathbf{v_2}$. De este modo, obtenemos dos autovectores no normalizados: $\mathbf{v_1} = (5, -2 + \sqrt{29})'$ y $\mathbf{v_2} = (2 - \sqrt{29}, 5)'$. Es inmediato comprobar que estos dos vectores son ortogonales.

Los autovectores normalizados son: $\mathbf{e_1} = (0{,}83, 0{,}56)'$ y $\mathbf{e_2} = (-0{,}56, 0{,}83)'$. Por lo tanto las componentes principales son:

$$Z_1 = 0{,}83X_1 + 0{,}56X_2$$
$$Z_2 = -0{,}56X_1 + 0{,}83X_2$$

Por último, calcularemos la proporción de varianza explicada por cada una de las dos componentes principales.

- La varianza explicada por la primera componente principal es: $\dfrac{\lambda_1}{\lambda_1 + \lambda_2} = \dfrac{6 + \sqrt{29}}{6 + \sqrt{29} + 6 - \sqrt{29}} = 0{,}95$.
- La varianza explicada por la segunda componente principal es: $\dfrac{\lambda_2}{\lambda_1 + \lambda_2} = \dfrac{6 - \sqrt{29}}{6 + \sqrt{29} + 6 - \sqrt{29}} = 0{,}05$.

2.2. Análisis de componentes principales dinámicas

El análisis de componentes principales (PCA) se basa en combinaciones lineales de las variables y esto no es lo más adecuado en procesos dinámicos. Por ese motivo, al trabajar con un conjunto de series temporales, $\{z_1, \ldots, z_T\}$, el uso del PCA para reducir la dimensión no es lo mejor, pues asume que las variables están incorreladas en el tiempo. Para solucionar este problema se van a definir, en esta sección, las componentes principales dinámicas.

Uno de los primeros en dar una solución a este problema fue Brillinger, quien lo resolvió de la siguiente manera: supongamos un proceso estacionario $\{z_t\}$, $-\infty < t < \infty$, con vector de media, m-dimensional, cero. La primera componente principal se define mediante la búsqueda de un vector $\mathbf{c_h}$, $-\infty < h < \infty$ y otro vector $\boldsymbol{\beta_j}$, $-\infty < j < \infty$, ambos de dimensión $m \times 1$, tal que si consideramos como primera componente principal dinámica la combinación lineal:

$$f_t = \sum_{h=-\infty}^{\infty} c_h' z_{t-h}$$

entonces

$$\mathbb{E}\left[\left(z_t - \sum_{j=-\infty}^{\infty} \beta_j f_{t+j}\right)' \left(z_t - \textstyle\sum_{j=-\infty}^{\infty} \beta_j f_{t+j}\right)\right]$$

es mínimo.
La resolución de este problema se puede consultar en Brillinger (2001) [3].

Para el cálculo de las componentes principales dinámicas, hemos utilizado el paquete *gdpc* de R creado por Daniel Peña, Ezequiel Smucler y Victor Yohai (para tener más información sobre este paquete puede consultarse [4]). El proceso para la obtención de las componentes principales dinámicas propuesto por estos autores presenta ciertas ventajas respecto al de Brillinger:

- Las componentes principales dinámicas no necesitan ser una combinación lineal de las observaciones.
- Las componentes principales dinámicas pueden estar basadas en una variedad de funciones de pérdida incluyendo las robustas.
- En el proceso de la obtención de las componentes principales dinámicas, las series pueden ser no estacionarias.

En este apartado explicaremos las componentes principales dinámicas basándonos en el artículo de Peña y Yohai (2016) [12].

Definición 2.2.1. *Supongamos que observamos m series temporales $Z_{j,t}$, $1 \leq j \leq m$, $1 \leq t \leq T$, y consideremos dos números enteros $k_1 \geq 0$ y $k_2 \geq 0$. Definimos las* ***primera componente principal dinámica****, con k_1 retardos y k_2 adelantos, como un vector $\boldsymbol{f} = f_t$, $-k_1 + 1 \leq t \leq T + k_2$.*

Intentemos reconstruir las series $Z_{j,t}$, $1 \leq j \leq m$. Para ello, podemos suponer que dicha reconstrucción se puede obtener como una combinación lineal de $\mathbf{f}$, siendo está óptima mediante el criterio de error cuadrático medio (MSE). Es decir, dado un posible factor $\mathbf{f}$, una matriz $\boldsymbol{\gamma} = \gamma_{j,i}^*$ de dimensión $m \times (k_1 + k_2)$ con $1 \leq j \leq m$, $-k_1 \leq i \leq k_2$ y un vector $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_m)$, la reconstrucción de $Z_{j,t}$ es:

$$\widehat{Z}_{j,t} = \sum_{i=-k_1}^{k_2} \gamma_{j,i} f_{t+i} + \alpha_j$$

Dado $K = k_1 + k_2$, definimos:

$$f_t^* = f_{t-k_1},\ 1 \leq t \leq T + K \qquad \beta_{j,i} = \gamma_{j,i-1-k_1},\ 1 \leq i \leq K+1 \qquad \beta_{j,h}^* = \gamma_{j,h-k_1-1},\ 1 \leq h \leq K+1$$

$$f_t^{**} = f_{t+K}^*,\ 1 - K \leq t \leq T \qquad \beta_{j,h}^{**} = \beta_{j,K+2-h}^*,\ 1 \leq h \leq K+1$$

Entonces, la reconstrucción de la serie se puede obtener como:

$$\widehat{Z}_{j,t} = \sum_{i=1}^{K+1} \beta_{j,i} f_{t+i-k_1} + \alpha_j = \sum_{h=0}^{K} \beta_{j,h+1}^* f_{t+h+1}^* + \alpha_j = \sum_{h=0}^{K} \beta_{j,h+1}^{**} f_{t-h}^{**} + \alpha_j$$

Entonces podemos usar, indistintamente, K retardos o K adelantos de las componentes principales para reconstruir la serie. Aunque la reconstrucción de la serie con K retardos es la más intuitiva, se obtendrá la solución óptima usando K adelantos, pues si tratamos con retardos las ecuaciones serán más complicadas, pues tendríamos que tratar con subíndices negativos.

Definición 2.2.2. *Dado $\boldsymbol{f} = (f_1, \ldots, f_{T+K})'$, $\boldsymbol{\beta} = \beta_{j,i}$ con $1 \leq j \leq m$, $1 \leq i \leq K+1$ y $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_m)$, la función de error cuadrático medio cuando reconstruimos las m series utilizando K adelantos es:*

$$MSE(\boldsymbol{f}, \boldsymbol{\beta}, \boldsymbol{\alpha}) = \frac{1}{Tm} \sum_{j=1}^{m} \sum_{t=1}^{T} (Z_{j,t} - \widehat{Z}_{j,t}(\boldsymbol{f}, \beta_j, \alpha_j))^2$$

Nótese que la función anterior está bien definida incluso cuando las series de tiempo no sean estacionarias.

Una elección óptima de $\boldsymbol{f} = (f_1, \ldots, f_{T+K})'$, $\boldsymbol{\beta} = \beta_{j,i}$ *con* $1 \leq j \leq m$, $1 \leq i \leq K+1$ *y* $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_m)$ *es aquella que verifica:*

$$(\widehat{\boldsymbol{f}}, \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\alpha}}) = \arg \min_{\boldsymbol{f} \in \mathbb{R}, \boldsymbol{\beta} \in \mathbb{R}^{m \times (k+1)}, \boldsymbol{\alpha} \in \mathbb{R}^m} MSE(\boldsymbol{f}, \beta, \boldsymbol{\alpha})$$

Si $\mathbf{f}$ es una solución óptima, entonces $\gamma \mathbf{f} + \delta$ también sería una solución óptima, por lo que tendríamos infinitas soluciones. Así que se elige un $\mathbf{f}$ tal que $\sum_{t=1}^{T+K} f_t = 0$ y $\frac{1}{T+K} \sum_{t=1}^{T+K} f_t^2 = 1$.

Definición 2.2.3. *Se define* $\widehat{\boldsymbol{f}}$ *como la* ***primera componente principal dinámica*** *de orden K de las series observadas* $\{Z_{1,t}\} \ldots, \{Z_{m,t}\}$*, siendo la primera componente principal dinámica de orden cero la primera componente principal clásica de los datos.*

Definición 2.2.4. *Se define la* ***segunda componente principal dinámica*** *como la primera componente dinámica de los residuos. Órdenes más altos de las componentes principales se definen de una manera similar.*

Una vez calculadas las componentes principales, tendremos que elegir el número adecuado de componentes para reducir la dimensión. Para ello, utilizaremos el criterio de la varianza explicada, definida en el apartado 2.1.3.

Capítulo 3

Metodología Box-Jenkins

A lo largo de este capítulo, se va a considerar un conjunto de variables aleatorias Z_t donde t toma valores en un conjunto C, siendo este un conjunto ordenado según los instantes temporales (días, meses, trimestres, años, ...). Entonces, para cada valor t del conjunto C tendremos una variable aleatoria, Z_t, y el conjunto de los valores observados en cada instante formarán una serie temporal, es decir, tendremos un vector de T variables aleatorias ordenadas en le tiempo, $(Z_1, Z_2, \ldots, Z_t, \ldots, Z_T)$, a la que llamaremos serie temporal.

Las series de tiempo tiene una estructura de dependencia, la cual trataremos de modelizar utilizando, para ello, la metodología Box-Jenkins, ver Peña (2010) [11]. Con dicha metodología trataremos de construir un modelo que sea capaz de captar la dinámica de la serie y así poder efectuar predicciones de valores futuros de dicha serie. Los pasos a seguir son:

- Identificar el modelo.
- Estimar los parámetros del modelo.
- Chequear el modelo.
- Predecir valores futuros basándonos en el modelo ajustado.

Este capítulo se va a organizar de la siguiente manera: en la sección 3.1 se estudiarán los modelos para procesos estacionarios, los cuales incluyen los modelo $AR(p)$, $MA(q)$ y $ARMA(p, q)$, tanto con período estacional como sin él; en la sección 3.2 se estudiarán los modelos para procesos no estacionarios; en la sección 3.3 se obtendrán estimadores de los parámetros de los modelos, tanto por mínimos cuadrados como por máxima verosimilitud; en la sección 3.4 se expondrán técnicas para chequear el modelo; en la sección 3.5 veremos cómo seleccionar, entre los modelos válidos, el más adecuado, y por último; en la sección 3.6 se estudiará cómo obtener predicciones de valores futuros a partir del modelo seleccionado.

3.1. Modelos para procesos estacionarios

Una serie temporal es estacionaría en *sentido estricto* si:

- Las distribuciones marginales de todas las variables son idénticas (como consecuencia la media y la varianza son la misma).
- las distribuciones finito-dimensionales de cualquier conjunto de variables sólo dependen de los retardos entre ellos.

La estacionariedad en *sentido estricto* es difícil de contrastar ya que sería necesario disponer de las distribuciones conjuntas para cualquier selección de variables del proceso. Por este motivo se define la estacionariedad en *sentido débil*. Diremos que un proceso es estacionario en sentido débil si:

- $\mu_t = \mu = cte.$

- $\sigma_t = \sigma = cte.$
- $\gamma(t, t-k) = \mathbb{E}\left[(z_t - \mu)(z_{t-k} - \mu)\right] = \gamma_k$, $k = 0, \pm 1, \pm 2, \ldots$, es decir, la covarianza entre dos variables depende únicamente de su separación.

Si una serie de tiempo pasa los contrastes de estacionariedad en *sentido débil* diremos que es una serie estacional.

Para modelar series estacionarias estudiaremos tres modelo: el autorregresivo, el de medias móviles y el ARMA.

3.1.1. Proceso autorregresivo de orden p, $AR(p)$

Una serie $\{Z_t\}$ sigue un proceso autorregresivo de orden p si:

$$Z_t = c + \phi_1 Z_{t-1} + \phi_2 Z_{t-2} + \cdots + \phi_{p-1} Z_{t-(p-1)} + \phi_p Z_{t-p} + a_t$$

siendo $c, \phi_1, \ldots, \phi_p$ constantes y $\{a_t\}$ un proceso de ruido blanco, conocido también con el nombre de innovaciones, con media cero y varianza σ_a^2.

Entonces, un proceso auntorregresivo de orden p, $AR(p)$, es aquel cuyo valor actual depende de los p instantes anteriores.

La ecuación anterior se puede reescribir como::

$$(1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p) Z_t - c = a_t$$

donde $\phi_p B^p Z_t = \phi_p Z_{t-p}$, siendo B el operador retardo ($BZ_t = B(Z_t) = Z_{t-1}$) y denotando B^p la composición de este operador consigo mismo p veces.

Si definimos $\phi_p(B) = 1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p$, el proceso autorregresivo de orden p, $AR(p)$, se puede expresar como:

$$\phi_p(B) Z_t - c = a_t$$

Nótese que para que proceso $AR(p)$ sea estacionario ha de ocurrir que todas las raíces del polinomio $\phi_p(B)$ sean de módulo mayor que 1.

Para más información sobre los $AR(p)$ puede consultarse Peña (2010) [11].

3.1.2. Proceso de medias móviles de orden q, $MA(q)$

Una serie $\{Z_t\}$ sigue un proceso de medias móviles de orden q si:

$$Z_t = c + a_t + \theta_1 a_{t-1} + \theta_2 a_{t-2} + \cdots + \theta_{q-1} a_{t-(q-1)} + \theta_q a_{t-q}$$

siendo $c, \theta_1, \ldots, \theta_q$ constantes y $a_t, a_{t-1}, \ldots, a_{t-q}$ innovaciones, con media cero y varianza σ_a^2.

Entonces, un proceso de medias móviles de orden q, $MA(q)$, es aquel cuyo valor actual no sólo depende de la última innovación sino que depende de las q últimas innovaciones.

La ecuación anterior se puede reescribir como:

$$Z_t = (1 - \theta_1 B - \theta_2 B^2 - \cdots - \theta_{q-1} B^{q-1} - \theta_q B^q) a_t - c$$

donde $\theta_q B^q a_t = \theta_q a_{t-q}$, siendo B el operador retardo ($Ba_t = B(a_t) = a_{t-1}$) y denotando por B^q a la composición de este operador consigo mismo q veces.

Si definimos $\theta_q(B) = 1 - \theta_1 B - \theta_2 B^2 - \cdots - \theta_{q-1} B^{q-1} - \theta_q B^q$, el proceso de medias móviles de orden q, $MA(q)$, se puede expresar como:

$$Z_t = \theta_q(B) a_t - c$$

Para que un proceso $MA(q)$ sea estacionario ha de ocurrir que todas las raíces del polinomio $\theta_q(B)$ sean de módulo menor que 1.

Para más información sobre los $MA(q)$ puede consultarse Peña (2010) [11].

3.1.3. Procesos $ARMA(p, q)$

Los procesos $ARMA(p, q)$ surgen como combinación de un proceso $AR(p)$ y un proceso $MA(q)$, es decir, los procesos $ARMA(p, q)$ son el resultado de añadir estructura MA a un AR, o lo contrario. Entonces, una serie $\{Z_t\}$ sigue un proceso $ARMA(p, q)$ si:

$$Z_t = c + \phi_1 Z_{t-1} + \phi_2 Z_{t-2} + \cdots + \phi_{p-1} Z_{t-(p-1)} + \phi_p Z_{t-p} + a_t + \theta_1 a_{t-1} + \theta_2 a_{t-2} + \ldots + \theta_{q-1} a_{t-(q-1)} + \theta_q a_{t-q}$$

siendo $c, \phi_1, \ldots, \phi_p, \theta_1, \ldots, \theta_q$ constantes y $a_t, a_{t-1}, \ldots, a_{t-q}$ innovaciones. Nótese que un $ARMA(p, 0) = AR(p)$ y un $ARMA(0, q) = MA(q)$.

Teniendo en cuenta la notación empleada en los apartados 3.1.1 y 3.1.2, los procesos ARMA(p,q) se pueden reescribir como:

$$\phi_p(B) Z_t - c = \theta_q(B) a_t$$

Para más información sobre los procesos ARMA(p,q) puede consultarse Peña (2010) [11].

3.1.4. Identificación de los procesos

En este apartado se explicará el proceso de identificación de cada uno de los modelos anteriores. Para ello será necesario definir los conceptos de: media muestral, función de autocovarianzas muestrales, función de autocorrelaciones simples muestrales y función de autocorrelaciones parciales muestrales.

Definición 3.1.1. *Sea $(Z_1, Z_2, \ldots, Z_T)$ una serie temporal. Se define la **media muestral** como:*

$$\overline{Z} = \frac{1}{T} \sum_{t=1}^{T} Z_t$$

Definición 3.1.2. *Sea $(Z_1, Z_2, \ldots, Z_T)$ una serie temporal. Se define la **función de autocovarianzas muestrales** entre observaciones separadas por k instantes como:*

$$\widehat{\gamma}_k = \frac{1}{T} \sum_{t=1}^{T-k} (Z_t - \overline{Z})(Z_{t-k} - \overline{Z})$$

con $\widehat{\gamma}_{-k} = \widehat{\gamma}_k$ para $k = 0, 1, 2, \ldots, T-1$.

Definición 3.1.3. *Sea $(Z_1, Z_2, \ldots, Z_T)$ una serie temporal. Se define la **función de autocorrelaciones simple muestrales** (fas) separada por k instantes como:*

$$\widehat{\rho}_k = \frac{\widehat{\gamma}_k}{\widehat{\gamma}_0}$$

Definición 3.1.4. *Sea $(Z_1, Z_2, \ldots, Z_T)$ una serie temporal. Se define la **función de autocorrelaciones parciales muestrales** (fap) como $\widehat{\alpha}_k = \widehat{\alpha}_{kk}$, siendo α_{kk} el estimador mínimo cuadrático de α_{kk} en la regresión*

$$Z_t = \alpha_{k0} + \alpha_{k1} Z_{t-1} + \cdots + \alpha_{kk} Z_{t-k} + \epsilon.$$

La identificación de las órdenes de los procesos definidos en los apartados 3.1.1, 3.1.2 y 3.1.3 se realiza observando la gráfica de la fas y fap de los residuos del ajuste. Para ello se tendrá en cuenta el cuadro 3.1.

	fas	fap
AR(p)	Muchos retardos no nulos	Primeros p-retardos no nulos, el resto 0
MA(q)	Primeros q-retardos no nulos, el resto 0	Muchos retardos no nulos
ARMA(p,q)	Muchos retardos no nulos	Muchos retardos no nulos

Cuadro 3.1: Identificación de los órdenes p y q

Para comprender mejor este proceso de identificación de los órdenes se realizarán algunos ejemplos. Para ello se simularán tres series e intentaremos observar el proceso mediante el cual se han generado.

En la figura 3.1 se muestra la primera serie simulada y las gráficas de las fas y fap. Claramente, se puede observar que la serie es estacionaria y que la fas tiene más estructura que la fap, pues en las fas hay muchos retardos no nulos y en la fap el retardo dos es el último en salirse. Teniendo en cuenta el cuadro 3.1, que la serie se ha generado mediante un proceso $AR(2)$.

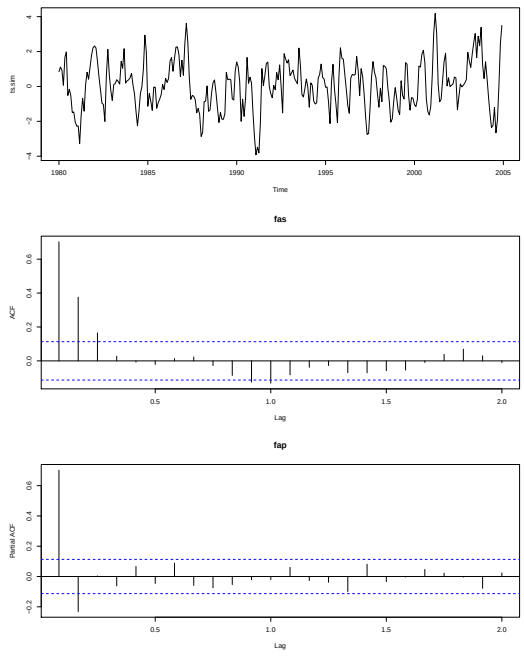


Figura 3.1: Serie, fas y fap

En la figura 3.2 se muestran la gráfica de la segunda serie simulada y las gráficas de las fas y fap. Se puede observar que la serie es estacionaria y que la fap tiene más estructura que la fas, pues en las fap hay muchos retardos no nulos y en la fas el retardo tres es el último en salirse. Teniendo en cuenta el cuadro 3.1, deducimos que la serie se ha generado mediante un proceso $MA(3)$.

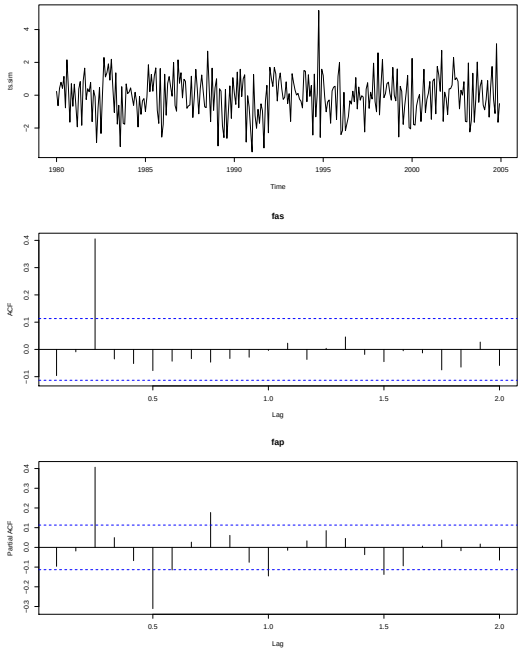


Figura 3.2: Serie, fas y fap

En la figura 3.3 se muestran la gráfica de la tercera serie simulada y las gráficas de las fas y fap. Se puede observar que la serie es estacionaria y que tanto el fas como el fap tienen mucha estructura. Esto nos indica que la serie ha sido generada mediante un proceso $ARMA(p, q)$. En este caso, se ha generado mediante un proceso $ARMA(2, 3)$.

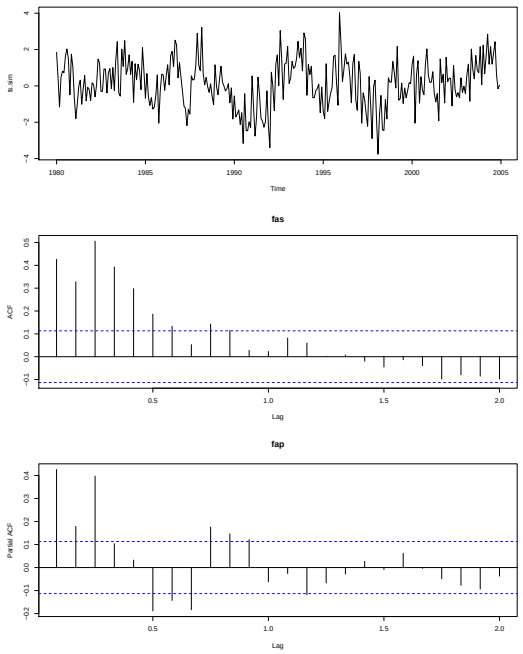


Figura 3.3: Serie, fas y fap

3.1.5. Modelos AR, MA y ARMA con período estacional

En algunas ocasiones lo esperado es que haya una dependencia entre observaciones separadas por un múltiplo de períodos estacionales, al que denominaremos s.

Una solución para este tipo de problemas es aumentar los ordenes p y q, lo que implicaría la introducción de muchos parámetros en el modelo. Por este motivo, se modificarán los modelos descritos en los apartados 3.1.1, 3.1.2 y 3.1.3 añadiéndoles una dependencia estacional.

Definición 3.1.5. *Una serie $\{Z_t\}$ sigue un proceso ARMA estacionario de órdenes estacionales autorregresivo P y de media móvil Q y período estacional s, $ARMA(P,Q)_s$, si:*

$$Z_t = c + \Phi_1 Z_{t-s} + \Phi_2 Z_{t-2s} + \cdots + \Phi_P Z_{t-Ps} + a_t + \Theta_1 a_{t-s} + \Theta_2 a_{t-2s} + \cdots + \Theta_Q a_{t-Qs}$$

Escrito de forma compacta:

$$\Phi(B^s) Z_t = c + \Theta(B^s) a_t$$

siendo:

$$\Phi(B^s) = 1 - \Phi_1 B^s - \Phi_2 B^{2s} - \cdots - \Phi_P B^{Ps} \qquad \Theta(B^s) = 1 + \Theta_1 B^s + \Theta_2 B^{2s} + \cdots + \Theta_Q B^{Qs}$$

y

$$\Phi_P B^{Ps} z_t = \Phi_P z_{t-Ps} \qquad \Theta_Q B^{Qs} z_t = \Theta_Q z_{t-Qs}$$

Nótese que:

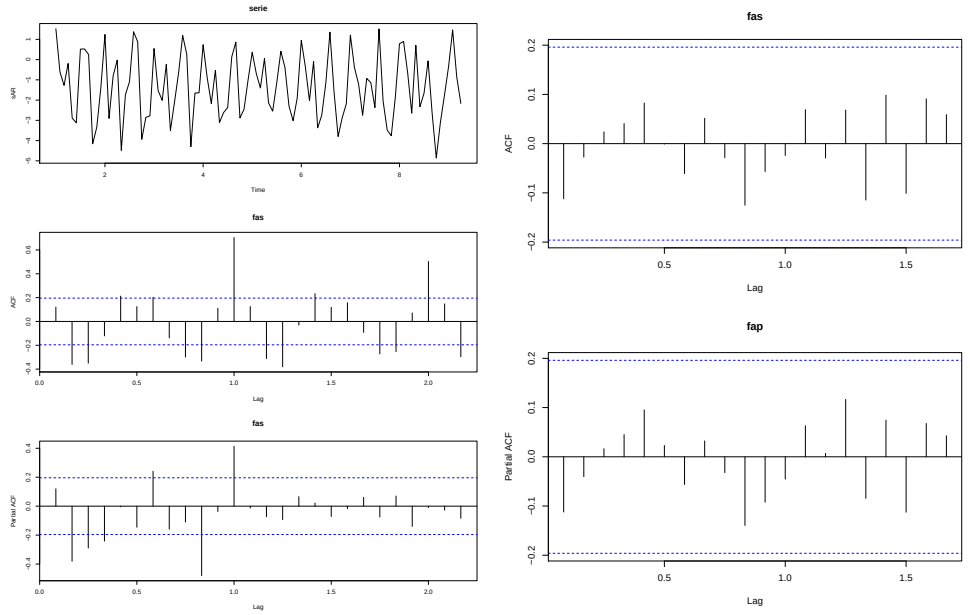
- Si $Q = 0$ estaríamos ante un proceso $ARMA(P,0)_s = AR(P)_s$, es decir, estaríamos ante un proceso autorregresivo de orden P y período estacional s.
- Si $P = 0$ estaríamos ante un proceso $ARMA(0,Q)_s = MA(Q)_s$, es decir, estaríamos ante un proceso de medias móviles de orden Q y período estacional s.

La identificación de los órdenes P y Q y del período estacional s se realiza observando la gráfica de la fas y de la fap. Para ello se debe tener en cuenta el cuadro 3.2.

	fas	fap
$AR(P)_s$	Muchos retardos $(s, 2s, \dots)$ no nulos	Se anula para los retardos mayores que Ps
$MA(Q)_s$	Se anula para los retardos mayores que Qs	Muchos retardos $(s, 2s, \dots)$ no nulos
$ARMA(P,Q)_s$	Muchos retardos $(s, 2s, \dots)$ no nulos	Muchos retardos $(s, 2s, \dots)$ no nulos

Cuadro 3.2: Identificación de los órdenes P y Q y el período estacional s

En la figura 3.4a se muestran la gráfica de la serie, las gráficas de las fas y fap. Se puede observar que la serie es estacional, que en la gráfica de las fas los retardos 12 y 24 no son nulos y que en la fap el retardo 12 es el único no nulo mientras que los restantes múltiplos de este período si lo son. Por lo tanto, la serie ha sido generada mediante un proceso $AR(1)_{12}$, pues al modelizar la serie mediante el proceso mencionado, la gráfica de la fas y fap de los residuos, figura 3.4b, nos muestran que estos son ruido blanco y además hemos comprobado que el modelo es válido.



(a) serie, fas y fap de la serie

(b) fas y fap de los residuos tras modelizar la serie mediante un proceso $AR(1)_{12}$

Figura 3.4: Proceso con período estacional

Para más información sobre estos modelos puede consultarse Shumway y Stoffer (2006) [13].

3.2. Modelos para procesos no estacionarios

En la práctica, lo más habitual es que las series no sean estacionarias, sino que presenten heterocedasticidad, tendencia y/o componente estacional.

La **heterocedasticidad** nos indica que la variabilidad de la serie no es constante, es decir, que aumenta o disminuye al hacerlo en nivel de la serie. Esto puede observarse en el gráfico de la misma. Para solucionar este problema se transforma la serie aplicándole un logaritmo neperiano, siendo esto lo más habitual, o realizando una transformación Box-Cox. La familia de transformaciones Box-Cox se define como aquella que transforma Z_t en:

$$\begin{cases} \dfrac{Z_t^\lambda - 1}{\lambda} & \text{si } \lambda \neq 0 \\ \log(Z_t) & \text{si } \lambda = 0 \end{cases}$$

Para más información sobre esta transformación se puede consultar Box and Cox (1964) [2].

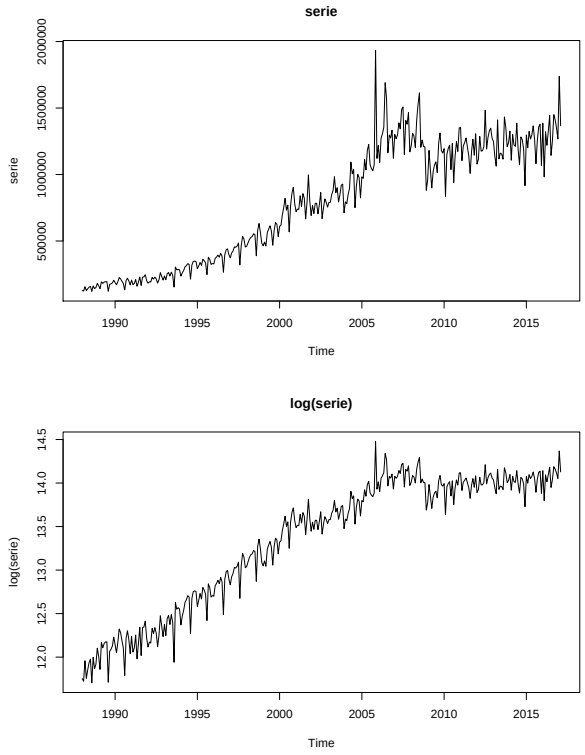


Figura 3.5: Gráfico de la serie frente al tiempo (arriba) y del logaritmo de la serie frente al tiempo (abajo)

En la figura 3.5 se puede observar que la serie presenta homocedasticidad, mientras que al aplicarle el logaritmo a dicha serie parece que la variabilidad se estabiliza. Sin embargo, aún estabilizando la variabilidad la serie no es estacional, pues presenta tendencia.

La **tendencia** se puede observar en el gráfico de la serie, pues muestra que la serie tiene una cierta "pendiente" positiva o negativa. También se puede observar en la fas muestral, en la cual los retardos toman valores positivos, estando a menudo los primeros de ellos cerca de 1 y después decaen lentamente a cero a medida que el retardo crece. Para eliminar la tendencia se le aplica a la serie sucesivas diferenciaciones regulares, denominadas d. Si después de diferenciar regularmente la serie no es estacionaria se vuelve a diferenciar la serie hasta que esta sea estacionaria. En general, se suele diferenciar como mucho 3 veces, $d \leq 3$. Una vez que es estacionaria se puede modelizar mediante un ARMA(p,q). A este proceso de le denomina ARIMA(p,d,q).

Definición 3.2.1. *Un proceso ARIMA(p,d,q) es aquel, que después de aplicarle d diferenciaciones regulares a la serie, se convierte en un ARMA(p,q), es decir, Z_t es un ARIMA(p,d,q) si el proceso $(1-B)^d Z_t$ es un ARMA(p,q).*

Recordando la notación del apartado 3.1.3, diremos que Z_t es un proceso ARIMA(p,d,q) si admite una representación del tipo:

$$\phi(B)(1-B)^d z_t = c + \theta(B)a_t$$

donde ni $\phi(B)$ ni $\theta(B)$ tienen raíces de módulo 1.

En la figura 3.6 se puede observar que la serie de la gráfica superior presenta tendencia y al aplicarle una diferenciación regular a dicha serie, gráfica inferior, la serie ya no muestra tendencia, sin embargo, se aprecia la presencia de componente estacional. Por lo que la serie aún no es estacionaria.

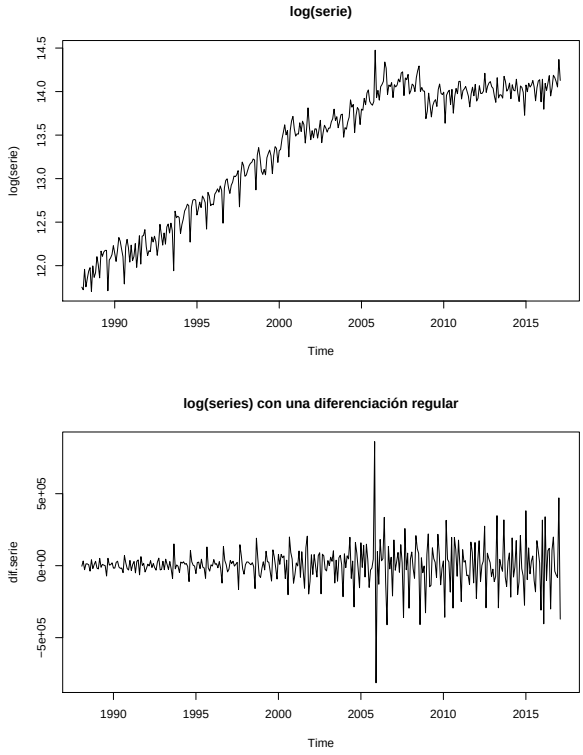


Figura 3.6: Gráficos de la serie sin diferenciar frente al tiempo (arriba) y de la serie diferenciada regularmente (abajo)

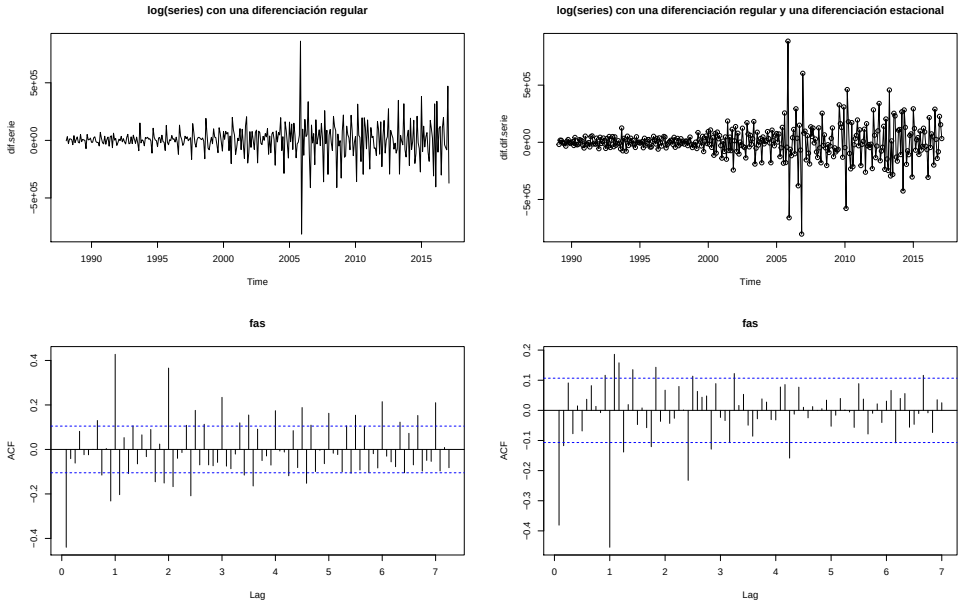
La **componente estacional** de una serie nos indica que la media de las observaciones no es constante sino que tiene un patrón cíclico. Se puede detectar en el gráfico secuencial de la serie y en la fas muestral. En este último se observa una fuerte correlación positiva en los retardo estacionales y, posiblemente, en sus múltiplos, convergiendo los retardos lentamente a cero a medida que crecen. Para eliminar la componente estacional se le aplica a la serie diferenciaciones estacionales hasta que la serie sea estacionaria. Al número de estas diferenciaciones se le denomina D y, normalmente, llega con aplicar una. Una vez que es estacionaria se puede modelizar mediante un $ARMA(p,q)$ (sólo dependencia regular), $ARMA(P,Q)_s$ (sólo dependencia estacional) o $ARMA(p,q) \times (P,Q)_s$ (ambos tipos de dependencia).

Definición 3.2.2. *Diremos que un $ARIMA(p,d,q) \times (P,D,Q)_s$ (o ARIMA estacional multiplicativo) es aquél que, después de aplicarle d diferencias regurales y D diferencias estacionales de período s, se convierte en un proceso $ARMA(p,q) \times (P,Q)_s$.*

Recordando la notación de los apartados 3.1.3 y 3.1.5, diremos que Z_t es un proceso $ARIMA(p,d,q) \times (P,D,Q)_s$ si admite una representación del tipo:

$$\phi(B)\Phi(B^s)(1-B)^d(1-B^s)^D Z_t = c + \theta(B)\Theta(B^s)a^t$$

En la figura 3.7a, tanto en el gráfico de la serie como en el de la fas, se puede observar la presencia de componente estacional, en este caso $s = 12$, pues en la fas los retardos que se salen son los múltiplos de 12. Aplicándole una diferenciación estacional a dicha serie, figura 3.7b, resulta ya estacional.



(a) Serie con una diferenciación regular

(b) Serie con una diferenciación regular y una diferenciación estacional

Figura 3.7: Serie sin diferencias estacionales (izquierda) y serie con diferencias estacionales (derecha)

Para más información sobre los procesos $ARIMA(p,d,q)$, $ARIMA(p,d,q)\times(P,D,Q)_s$ y las diferenciaciones se puede consultar Shumway y Stoffer (2006) [13]

3.3. Estimaciones de los parámetros

Supongamos que la serie de tiempo, $Z_1, Z_2, \ldots, Z_T$, ha sido generada por un proceso $ARMA(p,q)$. Recordemos que entonces dicha serie viene modelizada por:

$$Z_t = c + \phi_1 Z_{t-1} + \phi_2 Z_{t-2} + \cdots + \phi_p Z_{t-p} + a_t + \theta_1 a_{t-1} + \theta_2 a_{t-2} + \cdots + \theta_q a_{t-q}$$

donde los parámetros $c, \phi_1, \phi_2, \ldots, \phi_p, \theta_1, \theta_2, \ldots, \theta_q, \sigma_a$ son desconocidos. En esta sección se abordará el problema de estimación de los parámetros. Para ello se realizará una estimación por mínimos cuadrados y máxima verosimilitud, siendo esta última la que se utilizará en la práctica.

3.3.1. Estimación por mínimos cuadrados

Se definen los residuos de las estimaciones, $\tilde{c}, \tilde{\phi}_1, \tilde{\phi}_2, \ldots, \tilde{\phi}_p, \tilde{\theta}_1, \tilde{\theta}_2, \ldots, \tilde{\theta}_q$, para $t = 1, \ldots, T$, como la diferencia entre los valores observados y sus correspondientes estimaciones. Entonces:

$$\widehat{a}_t = Z_t - (\tilde{c} + \tilde{\phi}_1 Z_{t-1} + \tilde{\phi}_2 Z_{t-2} + \cdots + \tilde{\phi}_p Z_{t-p} + \tilde{\theta}_1 \widehat{a}_{t-1} + \tilde{\theta}_2 \widehat{a}_{t-2} + \cdots + \tilde{\theta}_q \widehat{a}_{t-q})$$

siendo la suma de los cuadrados:

$$S(\tilde{c}, \tilde{\phi}_1, \tilde{\phi}_2, \ldots, \tilde{\phi}_p, \tilde{\theta}_1, \tilde{\theta}_2, \ldots, \tilde{\theta}_q) = \sum_{t=1}^{T} \widehat{a}_t^2$$

La estimación de los parámetros, $c, \phi_1, \phi_2, \ldots, \phi_p, \theta_1, \theta_2, \ldots, \theta_q$, por el **método de mínimos cuadrados** se obtiene a través de los valores que minimizan la función S, es decir:

$$(\widehat{c}, \widehat{\phi}_1, \widehat{\phi}_2, \ldots, \widehat{\phi}_p, \widehat{\theta}_1, \widehat{\theta}_2, \ldots, \widehat{\theta}_q) = \arg \min_{\tilde{c}, \tilde{\phi}_1, \tilde{\phi}_2, \ldots, \tilde{\phi}_p, \tilde{\theta}_1, \tilde{\theta}_2, \ldots, \tilde{\theta}_q} S(\tilde{c}, \tilde{\phi}_1, \tilde{\phi}_2, \ldots, \tilde{\phi}_p, \tilde{\theta}_1, \tilde{\theta}_2, \ldots, \tilde{\theta}_q) = \arg \min \sum_{t=1}^{T} \widehat{a}_t^2$$

Si $p > 0$ se añade una dificultad a la estimación por mínimos cuadrados de los parámetros $\widehat{a}_1, \widehat{a}_1, \ldots, \widehat{a}_q$, pues dependen de valores no observados $Z_0, Z_{-1}, \ldots, Z_{1-p}$. Por lo que en vez de minimizar la función S, minimizaremos la función:

$$S_c(\tilde{c}, \tilde{\phi}_1, \tilde{\phi}_2, \ldots, \tilde{\phi}_p, \tilde{\theta}_1, \tilde{\theta}_2, \ldots, \tilde{\theta}_q) = \sum_{t=p+1}^{T} \widehat{a}_t^2$$

Pero $\widehat{a}_{p+1} = Z_t - (\tilde{c} + \tilde{\phi}_1 Z_p + \tilde{\phi}_2 Z_{p-1} + \cdots + \tilde{\phi}_p Z_1 + \tilde{\theta}_1 \widehat{a}_p + \tilde{\theta}_2 \widehat{a}_{p-1} + \cdots + \tilde{\theta}_q \widehat{a}_{p+1-q})$ depende de los valores $\widehat{a}_p, \widehat{a}_{p-1}, \ldots, \widehat{a}_{p+1-q}$ que a su vez depende de valores no observados de Z_t. Por lo que si se fijan los valores $\widehat{a}_p, \widehat{a}_{p-1}, \ldots, \widehat{a}_{p+1-q}$ se podrán reconstruir iterativamente los valores $\widehat{a}_{p+1}, \widehat{a}_{p+2}, \ldots, \widehat{a}_T$

Entonces la estimación de los parámetros, $c, \phi_1, \phi_2, \ldots, \phi_p, \theta_1, \theta_2, \ldots, \theta_q$, por el **método de mínimos cuadrados condicionados** se obtiene a través de:

$$\begin{cases} \text{mín } S_c(\tilde{c}, \tilde{\phi}_1, \tilde{\phi}_2, \ldots, \tilde{\phi}_p, \tilde{\theta}_1, \tilde{\theta}_2, \ldots, \tilde{\theta}_q) = \text{mín } \displaystyle\sum_{t=p+1}^{T} \widehat{a}_t^2 \\ \\ \text{s.a.} \qquad \widehat{a}_p = \widehat{a}_{p-1} = \cdots = \widehat{a}_{p+1-q} = 0 \end{cases}$$

3.3.2. Estimación por máxima verosimilitud

La estimación de máxima verosimilitud de los parámetros se obtiene a través de los valores que dan una "mayor credibilidad" y esta credibilidad se "mide" a través de la función de verosimilitud:

$$L_{Z_1, Z_2, \ldots, Z_T}(\tilde{c}, \tilde{\phi}_1, \ldots, \tilde{\phi}_p, \tilde{\theta}_1, \ldots, \tilde{\theta}_q, \tilde{\sigma}_a) = f_{\tilde{c}, \tilde{\phi}_1, \ldots, \tilde{\phi}_p, \tilde{\theta}_1, \ldots, \tilde{\theta}_q, \tilde{\sigma}_a}(Z_1, \ldots, Z_t)$$

siendo $f_{\tilde{c}, \tilde{\phi}_1, \ldots, \tilde{\phi}_p, \tilde{\theta}_1, \ldots, \tilde{\theta}_q, \tilde{\sigma}_a}$ la función de densidad conjunta de un vector aleatorio $(\tilde{Z}_1, \ldots, \tilde{Z}_t)'$ procedente de un proceso $ARMA$ con parámetros $\tilde{c}, \tilde{\phi}_1, \ldots, \tilde{\phi}_p, \tilde{\theta}_1, \ldots, \tilde{\theta}_q, \tilde{\sigma}_a$.

Entonces, la estimación de los parámetros $c, \phi_1, \ldots, \phi_p, \theta_1, \ldots, \phi_q, \sigma_a$ por medio del **método de máxima verosimilitud** se obtiene maximizando la función de verosimilitud L, es decir:

$$(\widehat{c}, \widehat{\phi}_1, \ldots, \widehat{\phi}_p, \widehat{\theta}_1, \ldots, \widehat{\theta}_q, \widehat{\sigma}_a) = \arg \max_{\tilde{c}, \tilde{\phi}_1, \ldots, \tilde{\phi}_p, \tilde{\theta}_1, \ldots, \tilde{\theta}_q, \tilde{\sigma}_a} L(\tilde{c}, \tilde{\phi}_1, \ldots, \tilde{\phi}_p, \tilde{\theta}_1, \ldots, \tilde{\theta}_q, \tilde{\sigma}_a)$$

Para más información sobre el método de máxima verosimilitud puede consultarse Peña (2010) [11].

3.4. Diagnosis del modelo

Una vez construido el modelo se comprobará si es válido. Para ello debe cumplir ciertas hipótesis. En el caso de que no las cumpla se tendrá que volver a ajustar un modelo hasta que este sea válido.

Una de las condiciones más importantes es que las innovaciones, a_t, sean ruido blanco, es decir:

1. $\mu_t = 0$ para $t = 1, \ldots, T$.
2. $\sigma_t^2 = \sigma_a^2$ para $t = 1, \ldots, T$.
3. $\gamma(s,t) = \mathbb{E}(a_s, a_t) = \begin{cases} \sigma_a^2 & \text{si } s = t \\ 0 & \text{si } s \neq t \end{cases}$
4. $a_t \in N(0, \sigma_a)$

La condición 1 nos índica que la esperanza es siempre constante e igual a cero, la condición 2 que la varianza es constante, la condición 3 que las variables están incorreladas y la condición 4 que las innovaciones siguen una distribución normal. Esta condición es importante ya que bajo normalidad la incorrelación implica independencia, por lo que no estaríamos dejando información por modelizar. Además, los estimadores serían asintóticamente eficientes y al obtener predicciones las podríamos acompañar con sus intervalos de confianza.

Para la diagnosis del modelo nos podremos ayudar de:

- **Gráficos**. La representación de la muestra frente al tiempo nos puede ayudar a detectar de manera visual la presencia de tendencia, componente estacional, variabilidad no constante y dependencia lineal. Cualquiera de estas situaciones invalidaría nuestro modelo.

  El gráfico Q-Q normal, que representa los cuantiles muestrales frente a los cuantiles de una distribución normal, nos indica si los residuos de nuestro modelo siguen una distribución normal. En caso de normalidad el gráfico debería ser aproximadamente lineal. En la figura 3.8 se puede observar como sería el Q-Q plot de una muestra que sigue una distribución normal (gráfico de la izquierda) y de una muestra que no sigue una distribución normal (gráfico de la derecha).

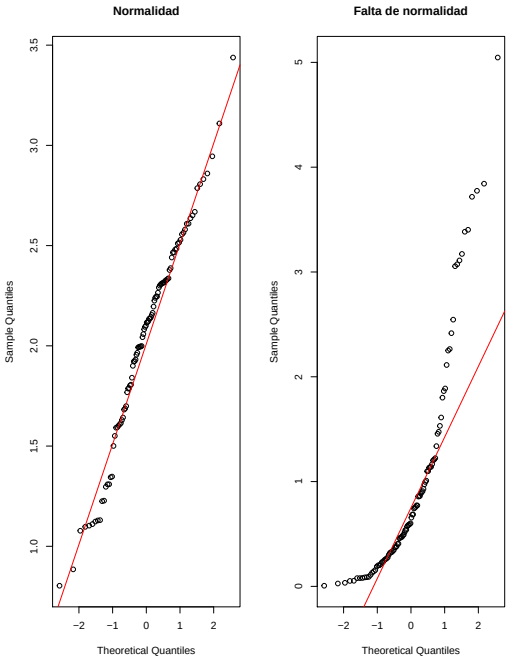


Figura 3.8: Q-Q plots

  Por último, el gráfico de los residuos nos permite saber si los residuos son homocedásticos y los gráficos de las fas y las fap nos indican si proceden de un proceso de ruido blanco.

- **Contraste de incorrelación**.

  En las autocorrelaciones simples y parciales se muestran las correlaciones de una en una, pudiendo ver si cada una de ellas es nula. Sin embargo, el **contraste de Ljung-Box** permite contrastar si los primeros h coeficientes (con h grande) son cero. En este contraste, la hipótesis nula es que las variables de la muestra sean incorreladas frente a la hipótesis alternativa de que sean correladas.

  El estadístico de contraste es:

$$Q(h) = T \times (T+2) \sum_{k=1}^{h} \frac{\hat{r}_k^2}{T-k}$$

que se distribuye, asintoticamente, como una χ^2 con grados de libertad igual a h si las correlaciones se estiman empíricamente de la serie original o igual a $h-p$, siendo p el número de parámetros, si las correlaciones se estiman empíricamente a partir de los residuos de una serie temporal previamente ajustada a un modelo, por ejemplo $ARMA$. Nótese que en la expresión $\hat{r}_k$ hace referencia a los residuos del modelo y h es el número de retardos que se están probando.

Para el contraste de Ljung-Box utilizaremos la función *Box.test(residuos, type="Ljung-Box")* de *R*, la cuál nos devuelve, entre otras cosas, el p-valor. Si el p-valor es menor que un cierto nivel de significación α, en nuestro caso tomaremos $\alpha = 0{,}05$, rechazaremos la hipótesis nula, en caso contrario, no habrá evidencias suficientes para rechazarla.

- **Contraste de media cero**.

  La hipótesis nula de este contraste es que los residuos del modelo provienen de variables aleatorias idénticamente distribuidas, con media cero y varianza finita. El estadístico de este contraste, asumiendo que T es grande, es:

$$\frac{\overline{\mathbf{r}}}{S_{\mathbf{r}}/\sqrt{T}}$$

  que se distribuye mediante una $N(0,1)$, donde $\mathbf{r} = (r1, \ldots, r_T)$ denota a los residuos del modelo.

  Por lo tanto, se rechaza la hipótesis nula con un nivel de significación α, en nuestro caso $\alpha = 0{,}05$, si:

$$|\overline{\mathbf{r}}| \geq z_{\alpha/2} \frac{S_{\mathbf{r}}}{\sqrt{T}}$$

- **Contraste de normalidad**.

  Denotemos por $G_1 = \dfrac{\sum_{t=1}^{T}(r_t - \overline{\mathbf{r}})^3}{\widehat{\sigma}^3}$ el coeficiente de asimetría y por $G_2 = \dfrac{\sum_{t=1}^{T}(r_t - \overline{\mathbf{r}})^4}{\widehat{\sigma}^4}$ el coeficiente de curtosis. La hipótesis nula de este contraste es que los residuos del modelo, $\mathbf{r} = (r_1, \ldots, r_T)$, provienen de variables aleatorias identicamente distribuidas con distribución gaussiana, frente a la hipótesis alternativa de que la distribución no sea gaussiana. El estadístico de contraste, asumiendo que T es grande, es:

$$\frac{TG_1^2}{6} + \frac{T(G_2-3)^2}{24}$$

  que sigue una distribución χ^2 con dos grados de libertad.

  Para el contraste de normalidad utilizaremos la función *jarque.bera.test(residuos)* de *R*, la cuál, entre otras cosas, devuelve un p-valor. Si el p-valor es menor que un nivel de significación α, en nuestro caso $\alpha = 0{,}05$, se rechazará la hipótesis nula, en caso contrario, no habrá evidencias suficientes para rechazarla.

3.5. Criterio selección del modelo

Lo habitual es que nos encontremos con un conjunto de modelos, $M_1, \ldots, M_m$, válidos. En este punto lo que desearíamos es seleccionar el que mejor explique la serie observada. En esta sección veremos distintos criterios para seleccionar el modelo más adecuado. Para ello, definiremos k como la cantidad de parámetros que contiene el proceso $ARMA(p,q)$, es decir, $k = p + q + 1$ o $k = p + q$ si tenemos en cuenta la constante o no, respectivamente, y por $\widehat{\Phi}_{k+1}$ el vector formado por las estimaciones de máxima verodimilitud de dichos coeficientes y de σ_a.

- **Criterio AIC de Akaike**. $AIC = -2\log(L(\widehat{\Phi}_{k+1})) + 2k$. Este criterio tiene un problema y es que tiende a sobrestimar el número de parámetros del modelo, teniendo un efecto más grande en muestras de tamaño pequeño.
- **Criterio AIC corregido**. $AICC = -2\log(L(\widehat{\Phi}_{k+1})) + \dfrac{2(kT+k+2)}{T-k-2}$. Este criterio surge como una alternativa al AIC para corregir la sobrestimación.
- **Criterio BIC**. $BIC = -2\log(L(\widehat{\Phi}_{k+1})) + k\log(T)$. Este criterio penaliza más la introducción de nuevos parámetros que el criterio AIC, con lo que tiende a elegir modelos más parsimoniosos. La diferencia entre ambos puede ser grande cuando T lo es.

Supongamos que tenemos varios modelos, pues seleccionaremos aquel que tenga un menor, por ejemplo, AIC. En el caso de que la diferencia de los AICs sea menor de dos unidades, seleccionaremos el modelo que tenga un menor número de parámetros o el que sea más fácil de explicar.

Para más información sobre los criterios definidos anteriormente se puede consultar Peña (2010) [11].

3.6. Predicciones del modelo

Una vez elegido el modelo más adecuado para la serie temporal podremos predecir valores futuros del proceso a h instantes de tiempo, esto es, predecir el valor de Z_{T+h}, utilizando los parámetros estimados como si fueran los verdaderos. Además si el modelo es el correcto, estos parámetros minimizan el error de predicción a cualquier horizonte, en caso contrario, esto no sería necesariamente así.

Dicha predicción se denomina predicción con origen en T y horizonte h y se denota por $\widehat{z}_T(h)$.

- Supongamos que la serie temporal, $z_1, z_2, \ldots, z_T$, ha sido generada mediante un proceso $AR(1)$, es decir:

$$Z_t = c + \phi_1 Z_{t-1} + a_t$$

y queremos predecir en origen T a horizonte 1. Entonces:

$$Z_{T+1} = c + \phi_1 Z_T + a_{T+1}$$

en la cual el valor de Z_T es conocido (z_T) y los valores de c, ϕ_1 y a_{T+1} son desconocidos y serán sustituidos por sus estimaciones. Nótese que la predicción de a_{T+1} a partir de la serie es su media, pero $\mathbb{E}(a_{T+1}) = 0$ ya que la serie no contiene información sobre a_{T+1}.

Por lo tanto:

$$\widehat{z}_T(1) = \widehat{c} + \widehat{\phi}_1 z_T$$

Supongamos que ahora queremos predecir en origen T a horizonte 2, entonces:

$$Z_{T+2} = c + \phi_1 Z_{T+1} + a_{T+2}$$

donde los valores de c, ϕ_1, a_{T+2} y Z_{T+1} son desconocidos. Los tres primeros serán sustituidos por sus estimaciones y el valor de Z_{T+1} será sustituido por su predicción, $\widehat{z}_T(1)$. Nótese que la predicción de a_{T+2} a partir de la serie es su media, pero $\mathbb{E}(a_{T+2}) = 0$ ya que la serie no contiene información sobre a_{T+2}.

Por lo tanto:

$$\widehat{z}_T(2) = \widehat{c} + \widehat{\phi}_1 \widehat{z}_T(1) = \widehat{c} + \widehat{\phi}_1(\widehat{c} + \widehat{\phi}_1 z_T) = \widehat{c}(1 + \widehat{\phi}_1) + \widehat{\phi}_1^2 z_T$$

En general, si la serie de tiempo ha sido generada por un proceso $AR(p)$, las predicciones en origen T a horizonte h serán:

$$\widehat{z}_T(h) = \widehat{c} + \widehat{\phi}_1 \widehat{z}_{T+h-1} + \widehat{\phi}_2 \widehat{z}_{T+h-2} + \cdots + \widehat{\phi}_p \widehat{z}_{T+h-p}$$

donde $\widehat{z}_t = z_t$ si $t \leq T$. En caso contrario se sustituyen por sus respectivas predicciones.

Nótese que las predicciones a largo plazo de valores futuros ($h \to \infty$) de un proceso $AR(p)$ coinciden con la media del proceso.

- Supongamos que la serie temporal, $z_1, z_2, \ldots, z_T$ ha sido generada por un proceso $MA(1)$, es decir:

$$Z_t = c + a_t + \theta_1 a_{t-1}$$

y queremos predecir a origen T con horizonte 1. Entonces:

$$Z_{T+1} = c + a_{T+1} + \theta_1 a_T$$

Los valores de c, a_{T+1}, θ_1 y a_T no los tenemos y serán sustituidos por sus estimaciones. La predicción de a_{T+1} a partir de la serie es su media, pero $\mathbb{E}(a_{T+1}) = 0$ ya que la serie no contiene información sobre a_{T+1} y la predicción de a_T no es inmediata, pues la serie sí contiene información acerca de a_T. Denotaremos la predicción de a_T por $\widehat{a}_T(0)$.

por lo tanto:

$$\widehat{z}_T(1) = \widehat{c} + \widehat{\theta}_1 \widehat{a}_T(0)$$

Supongamos que queremos predecir en origen T a horizonte 2, entonces:

$$Z_{T+2} = c + a_{T+2} + \theta_1 a_{T+1}$$

Los valores de c, a_{T+2}, θ_1 y a_{T+1} no los tenemos y serán sustituidos por sus estimaciones. Las predicciones de a_{T+1} y a_{T+2} a partir de la serie son sus medias respectivamente, pero $\mathbb{E}(a_{T+1}) = 0$ y $\mathbb{E}(a_{T+2}) = 0$ ya que la serie no contiene ni información de a_{T+1} ni de a_{T+2}.

Por lo tanto:

$$\widehat{z}_T(2) = \widehat{c}$$

En general, si la serie de tiempo ha sido generada por un $MA(q)$, las predicciones en origen T con horizonte h serán:

$$\widehat{z}_T(h) = \widehat{c} + \widehat{a}_{T+h} + \theta_1 \widehat{a}_{T+h-1} + \theta_2 \widehat{a}_{T+h-2} + \cdots + \theta_q \widehat{a}_{T+h-q}$$

donde $\widehat{a}_t = 0$ si $t > T$. En caso contrario, serán sustituidos por sus predicciones, pues la serie contiene información sobre ellos. Para obtener dichas predicciones existen distintas técnicas.

- **Utilizar la invertibilidad del proceso**.

$$Z_t = c + a_t + \pi_1 Z_{t-1} + \pi_2 Z_{t-2} + \cdots$$

Despejando a_t y sustituyendo t por T obtenemos:

$$a_T = Z_T - c - \pi_1 Z_{t-1} - \pi_2 Z_{t-2} - \cdots - \pi_{T-1} Z_1 - \pi_T Z_0 - \cdots$$

donde $Z_t = z_t$ si $1 \leq t \leq T$ y $Z_t = \mu$ si $t \leq 0$.

- **Fijar un valor inicial para $\boldsymbol{a_1}$**, generalmente 0, y predecir recursivamente los valores de a_t para $t = 2, \ldots, T$ a partir de la serie $z_1, \ldots, z_T$.
- Predecir a_T a través de una **combinación lineal** de $z_1, \ldots, z_T$.

- Combinando los dos procesos anteriores, $AR(p)$ y $MA(q)$, se pueden predecir valores futuros para procesos $ARMA(p, q)$.

Cuando los residuos pasen el contraste de normalidad se podrá construir un intervalo de predicción. Para más información sobre ellos, puede consultarse Peña (2010) [11].

Capítulo 4

Modelos de regresión

El objetivo principal de este trabajo es obtener predicciones del producto interior bruto (PIB) a partir de las variables internas de ABANCA. Por lo que después de reducir la dimensión de nuestras variables explicativas, será necesario realizar algún modelo de regresión. En nuestro caso, la variable respuesta será el PIB o el consumo en hogares gallegos y las variables explicativas las componentes principales dinámicas. Una vez construido el modelo deberemos chequearlo para ver si es válido, pues en caso contrario no se podrá realizar predicciones y deberemos seleccionar otro.

Para la obtención de dichas predicciones estudiaremos distintos modelos de regresión, como la regresión simple, regresión dinámica, los modelos lineales generalizados y los modelos aditivos generalizados.

4.1. Modelo de regresión lineal simple

La regresión se suele formalizar como la media de la variable respuesta condicionada al valor que tome la variable explicativa. Se trataría de la función:

$$m(x) = \mathbb{E}\left[Y/X = x\right] \text{ para cada } x \in X$$

Entonces la variable respuesta, Y, se puede descomponer como la función anterior más un error de media cero, es decir:

$$Y = m(X) + \varepsilon$$

siendo ε el error, el cual verifica $\mathbb{E}(\varepsilon/X = x) = 0$, X la variable explicativa e Y la variable respuesta.

En esta sección se va a suponer que tanto la variables X como la variable Y son univariantes.

4.1.1. Estimación de los parámetros

Recordemos que los parámetros β_0 y β_1 son desconocidos por lo que se estimarán de modo que los residuos sean los más pequeños posibles, es decir, los estimadores por mínimos cuadrados son $\widehat{\beta}_0$ y $\widehat{\beta}_1$ tales que:

$$\sum_{i=1}^{n}(Y_i - \widehat{\beta}_0 - \widehat{\beta}_1 X_i)^2 = \min_{\beta_0,\beta_1} \sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 X_i)^2$$

La minimización anterior se lleva a cabo realizando las derivadas parciales respecto de cada parámetro, igualándolas a 0 y despejando $\widehat{\beta}_0$ y $\widehat{\beta}_1$. Obteniendo así que:

$$\widehat{\beta}_0 = \overline{Y} - \frac{S_{XY}}{S_X^2}\overline{X} \in N\left(\beta_0, \sigma^2\left(\frac{1}{n} + \frac{\overline{X}^2}{nS_X^2}\right)\right) \qquad \text{y} \qquad \widehat{\beta}_1 = \frac{S_{XY}}{S_X^2} \in N\left(\beta_1, \frac{\sigma^2}{nS_X^2}\right)$$

siendo $\overline{Y}$ e $\overline{X}$ las medias de la variable respuesta y explicativa respectivamente, S_{XY} la covarianza

entre ambas variables y S_X^2 la varianza de la variable explicativa.

Para más información sobre la estimación de los parámetros puede consultarse el capítulo 1 de Wood (2006) [14].

4.1.2. Diagnosis del modelo

El modelo lineal debe cumplir las siguientes hipótesis:

- **Linealidad**. La función de regresión es una linea recta, por lo que se puede expresar como:

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

  donde β_0 y β_1 son parámetros desconocidos y ε es una variable aleatoria no observable y de media cero.

  Esta hipótesis consiste en suponer que cuando la variable explicativa, X, toma valor 0, la media de la variable respuesta, Y, es β_0 y que la variable respuesta crece una cantidad constante, β_1, cada vez que se incrementa la variable explicativa en una unidad.

  Para comprobar el cumplimiento de esta hipótesis se representa el diagrama de dispersión de la muestra junto a la recta de regresión y el diagrama de dispersión de los errores frente a la variable explicativa. Si observamos que los datos tiene una evolución lineal, dicha hipótesis se cumple.

- **Homocedasticidad**. La varianza del error debe ser la misma, independientemente del valor de la variable explicativa, es decir:

$$Var(\varepsilon / X = x) = \sigma^2 \ \forall x \in X$$

  Para comprobar que dicha hipótesis es cierta se representan los mismos diagramas que en el punto anterior. Si en ellos observamos que la variabilidad no es constante, es decir, que crece o disminuye al aumentar los valores de X, rechazaremos dicha hipótesis.

- **Normalidad**. Los residuos deben seguir una distribución normal, es decir, $\varepsilon \sim N(0, \sigma^2)$.

  Para comprobar el cumplimiento de esta hipótesis se realiza el test de Shapiro-Wilk, utilizando para ello la función *shapiro.test(residuos)* de *R*, la cual nos devuelve, entre otras cosas, el p-valor. Si el p-valor es menor que un cierto nivel de significación α, en nuestro caso tomaremos $\alpha = 0{,}05$, rechazaremos la hipótesis nula, en caso contrario, no habrá evidencias suficientes para rechazarla.

- **Independencia**. Las variables aleatorias que representan el error, $\varepsilon_1, \ldots, \varepsilon_n$, son mutuamente independientes.

  Para comprobar el cumplimiento de dicha hipótesis se utilizará el test de Ljung-Box explicado en el apartado 3.4.

4.1.3. Modelo lineal general

Supongamos ahora que nuestra variable explicativa no es univariante, es decir, que poseemos un conjunto de variables explicativas , $X_1, X_2 \cdots, X_{p-1}$. Para obtener un modelo lineal vamos a considerar una combinación lineal de las variables explicativas, entonces:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_{p-1} X_{p-1} + \varepsilon$$

donde Y es la variable respuesta, $X_1, \ldots, X_{p-1}$ las variables explicativas, $\beta_0, \beta_1, \ldots, \beta_{p-1}$ son los parámetros y ε el error.

Así que podremos reescribir la ecuación como:

$$Y_i = \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + \cdots + \beta_{p-1} X_{i,p-1} + \varepsilon_i$$

siendo Y_i el i-ésimo valor de la variable respuesta, $X_{i,1}, X_{i,2}, \ldots, X_{i,p-1}$ el i-ésimo valor de cada variable explicativa y ε_i el i-ésimo error.

Reescribiéndolo en forma vectorial tenemos:

$$\begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} 1 & X_{1,1} & X_{1,2} & \cdots & X_{1,p-1} \\ 1 & X_{2,1} & X_{2,2} & \cdots & X_{2,p-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n,1} & X_{n,2} & \cdots & X_{n,p-1} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

y dando nombres a estos vectores y matrices obtenemos:

$$Y = X\beta + \varepsilon$$

Este modelo se conoce como modelo de regresión lineal múltiple. En este caso, el error, ε, también debe cumplir las hipótesis de homocedasticidad, normalidad e independencia mencionadas anteriormente.

Nótese que en este caso los parámetros $\beta_0, \beta_1, \ldots, \beta_{p-1}$ son también desconocidos, por lo que deberemos estimarlos, de modo que minimicen el error cometido. Así que obtendremos la estimación del vector β mediante el método de mínimos cuadrados, es decir, seleccionaremos el β que hace mínima la expresión $\sum_{i=1}^{n}(Y_i - x_i\beta)^2$. De modo que algunos cálculos nos permiten obtener una expresión para el estimador de mínimos cuadrados:

$$\widehat{\beta} = (X'X)^{-1}X'Y$$

Para más información sobre el modelo de regresión lineal múltiple, puede consultarse el capítulo 1 de Wood (2006) [14].

4.1.4. Criterio de selección del modelo

En este caso, al igual que ocurría en el capítulo 3, puede ocurrir que haya varios modelos que pasen los contraste de hipótesis y que lo que se quiera es elegir uno de ellos. Para solucionar este problema, se define el coeficiente de determinación ajustado, el cual se denota por $R^2 ajustado$.

$$R^2 ajustado = 1 - \frac{RSS/(n-p)}{TSS/(n-1)}$$

donde $RSS = \sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2$, $TSS = \sum_{i=1}^{n}(Y_i - \overline{Y}_i)^2$ y p los grados de libertad de RSS, que coinciden con el número de parámetros del modelo. Este coeficiente toma valores entre cero y uno, y cuanto más cerca de uno esté más próximo a la recta ajustada estarán los datos. Por lo que se elegiría aquél modelo que tenga un mayor R^2.

4.2. Regresión dinámica

En la sección anterior, 4.1, hemos visto los modelo de regresión simple. Supongamos que queremos aplicar el modelo a un conjunto de datos, en el cual, tanto la variable respuesta como la variable explicativa son series temporales. Por lo que el modelo sería:

$$Y_t = \beta_0 + \beta_1 X_t + \varepsilon_t$$

siendo ε un proceso de ruido blanco con varianza σ_a^2. Al intentar aplicarlo a variables que presentan una dependencia temporal nos vamos a encontrar con tres problemas, pues el modelo de regresión lineal simple:

- supone que la relación entre dos variables es instantánea, cuando en realidad la situación puede ser más compleja, pues puede ocurrir que la variable respuesta en cierto instante t, Y_t, dependa de la variable explicativa retardada k instantes, X_{t-k}, o incluso que dependa de todas las variables $(X_t, X_{t-1}, \ldots, X_{t-k})$.

- supone que la relación va de la variable explicativa en un cierto instante t, X_t, a la variable respuesta en el mismo instante, Y_t, pero que la relación al contrario no influye. Eso no tiene porque ser así. Un ejemplo ilustrativo es el que se menciona en Peña (2010) [11], el cual dice: "la inversión en publicidad está relacionada con las ventas a futuro pero las ventas del futuro están relacionadas con la inversión en publicidad del pasado".

- supone que la parte no explicada del modelo es un conjunto de variables independientes. Esta hipótesis no suele verificarse en datos dinámicos

Para solucionar estos problemas se va a estudiar el modelo de regresión dinámico. Para ello se definirán unos conceptos que nos indicarán la existencia, o no, de relación lineal sobre las series de tiempo.

Definición 4.2.1. *Dado dos procesos $\{X_s\}$ e $\{Y_t\}$ se define la* ***función de covarianzas cruzadas*** *como:*

$$\gamma_{s,t}(X, Y) = cov(X_s, Y_t)$$

Definición 4.2.2. *Dado dos procesos $\{X_s\}$ e $\{Y_t\}$ se define la* ***función de correlaciones cruzadas*** *como:*

$$\rho_{s,t}(X, Y) = \frac{\gamma_{s,t}(X, Y)}{\sigma_{X_s}\sigma_{Y_t}}$$

Definición 4.2.3. *Dado dos procesos $\{X_s\}$ e $\{Y_t\}$ se dicen* ***conjuntamente estacionarios*** *si:*

- *Ambos son estacionarios.*

- *Las covarianzas cruzadas sólo dependen del retardo entre las variables, es decir, $\gamma_{t,t-k}(X, Y) = \gamma_{s,s-k}(X, Y)$, $\forall t, s, k$.*

Para la detección de relación lineal entre dos procesos, $\{X_t\}$ e $\{Y_t\}$, debemos de realizar el preblanqueo sobre los procesos (si no son procesos estacionarios deberemos diferenciar las series) para eliminar distorsiones entre ellas, pues en caso contrario podemos llegar a la conclusión de que están relacionadas cuando realmente no lo están. Para más información sobre el preblanqueo de la serie se puede consultar las páginas 100-101 de Arnau Gras J (2001) [1] y la página 537 de Peña (2010) [11].

En la figura 4.1 se representan las correlaciones cruzadas de las series preblanqueadas. En la cual, se observa la existencia de relación lineal, pues hay un retardo que se sale de las bandas.

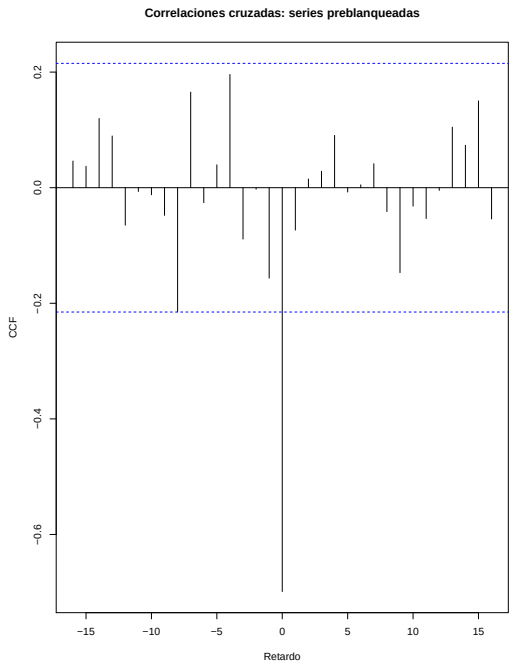


Figura 4.1: Correlaciones cruzadas

Como se ha mencionado anteriormente, en algunas ocasiones, la relación lineal entre los procesos no tiene porqué ser contemporánea. Pues, puede que la variable Y_t esté relacionada con la variable explicativa, X_t, retardada k instantes temporales, es decir, con X_{t-k}. Entonces:

$$Y_t = \beta_0 + \beta X_{t-k} + \varepsilon_t$$

donde $\varepsilon_t \in N(0, \sigma^2)$.

Una vez detectada la existencia de relación lineal y sugerido el retardo k, se procede a la estimación de los parámetros. Nótese que el parámetro k se selecciona observando la gráfica de las correlaciones cruzadas de las series preblanqueadas. Por ejemplo, observando la figura 4.1 se seleccionaría $k = 0$, pues es el retardo que se sale de las bandas. De modo, que se estaría ante un modelo de regresión lineal contemporáneo. Como se están considerando series temporales, lo más normal es que los errores, ε_t, estén correlados y/o los procesos, X_t y Y_t, no sean estacionarios. Entonces los pasos a seguir son:

1. Elegir el retardo temporal adecuado, observando para ello, el gráfico de las correlaciones cruzadas de las series preblanqueadas.
2. Una vez seleccionado el retardo, ajustar un modelo de regresión simple y observar los errores. Si los errores son estacionarios, pasar al siguiente paso, en caso contrario, diferenciar tanto la variable respuesta como la variable explicativa y volver a realizar este paso.
3. Proponer un modelo ARMA para los residuos de la regresión. Para ello, se va a tener cuenta lo visto en el capítulo 3.
4. Ajustar conjuntamente los parámetros de la regresión y del ARMA.
5. Chequear el modelo anterior, pues este debería de pasar los contrastes mencionados en 3.4. Si los errores del modelo no pasan alguno de los contrastes, se deberá volver al paso anterior.
6. Una vez que el modelo sea válido se podrán predecir valores futuros de manera análoga a lo explicado en la sección 3.6. Se tiene que tener en cuenta que, en algunos casos, se va a predecir el valor de la variable respuesta conociendo el valor que puede tomar la variable explicativa en ese instante del futuro.

4.3. Modelo lineal generalizado (GLM)

En los modelos de regresión simple la variable respuesta, Y, se asume Gaussiana y las covariables tienen un efecto lineal sobre la variable respuesta. Una extensión de este modelo es el **modelo lineal generalizado** (GLM), en el cual, la variable respuesta no se tiene porque asumir gaussiana, es decir, puede ser binaria, de recuento,...

Cuando la variable respuesta no sea Gaussiana, no va a ser posible una conexión directa entre el valor esperado y el predictor lineal, pues el dominio de $\mathbb{E}[Y/X_1,\dots,X_{p-1}]$ no es la recta real.

Recordemos que en un modelo de regresión lineal $\mathbb{E}[Y/X_1,\dots,X_{p-1}] = \beta_0+\beta_1X_1+\cdots+\beta_{p-1}X_{p-1} = \eta$. Por lo que en un GLM será necesario realizar una transformación, h, para expresar el dominio del valor esperado correctamente, es decr:

$$\mathbb{E}[Y/X_1,\dots,X_{p-1}] = h(\eta)$$

Si denotamos por $\mu = \mathbb{E}[Y/X_1,\dots,X_{p-1}]$, tenemos que $\mu = h(\eta) = h(\beta_0 + \beta_1X_1 + \cdots, \beta_{p-1}X_{p-1})$. Entonces:

$$g(\mu) = \eta = \beta_0 + \beta_1X_1 + \cdots, \beta_{p-1}X_{p-1} = \beta_0 + \sum_{j=1}^{p-1}\beta_jX_j$$

siendo $g = h^{-1}$ la función link o de enlace, η el predictor lineal y $\eta_j = \beta_jX_j$ el efecto parcial lineal de X_j, para $j = 1,\dots,p-1$.

En estos modelos, se debe elegir la función link y la distribución de la variable respuesta. Nótese que si se toma como función link la identidad y se asume que la variable respuesta es Gaussiana, se estará ante un modelo de regresión lineal simple.

4.3.1. Estimación de los parámetros

Los parámetros β son desconocidos y la estimación de ellos se obtiene maximizando la log-verosimilitud, $l(\beta)$, iterativamente mediante el algoritmo de Fisher scoring (IRLS):

$$\beta^{(k+1)} = (X^TW^{(k)}X)^{-1}X^TW^{(k)}z^{(k)}$$

siendo W una matriz diagonal de pesos y z la respuesta.

El parámetro final se denotará por $\widehat{\beta}$, que será aquel que minimice la deviance, $D(\beta) = \| \sqrt{W}(z - X\beta) \|^2$. La deviance es una cantidad que se interpreta de forma similar a la suma de residuos al cuadrado (RSS) de un modelo de regresión lineal simple. Viene dada por la expresión:

$$D = 2(l(\beta_{max}) - l(\widehat{\beta}))\phi$$

donde $l(\beta_{max})$ es la log-verosimilitud maximizada del modelo saturado, entendiendo por modelo saturado aquel que se ajusta perfectamente a los datos, y ϕ es el parámetro escala.

Para más información puede consultarse Müller (2004) [9] y el capítulo 2 de Wood (2017) [14].

4.3.2. Criterio de selección del modelo

Como ocurría en la sección anterior, en este caso, también puede ocurrir que haya varios modelos válidos. Para elegir uno de ellos se calcula el porcentaje de deviance explicada, el cual viene dado por:

$$\%\text{ Dev explicada} = \frac{Null\ deviance - Res\ deviance}{Null\ deviance} \times 100,$$

donde la *Null deviance* es la deviance del modelo que no depende de ninguna variable explicativa, es decir, el modelo que sólo incluye el intercepto para explicar el comportamiento de la variable respuesta, y la *Res deviance* es la diferencia entre la deviance del modelo que no depende de ninguna variable explicativa, *Null Deviance*, menos la deviance del modelo que incluye las variables explicativas consideradas. Nótese que *Null deviance − Res deviance = deviance del modelo que incluye variables explicativas*. Sin embargo, en la expresión se mantiene la diferencia, pues el summary del modelo nos devuelve esos dos valores.

Este concepto es análogo al R^2 en los modelos de regresión lineal simple.

4.3.3. Añadiendo flexibilidad al modelo

Aunque los modelos GLM ya son más flexibles que los modelos de regresión lineal, las covariables aún ejercen un efecto lineal sobre la variable respuesta, cuando puede que dichas covariables tengan un efecto no lineal desconocido, f, sobre la variable respuesta. En este caso, el modelo sería:

$$\eta = \beta_0 + f(X)$$

Para añadir esa flexibilidad se suele utilizar:

- La **regresión polinómica**, es decir, $\eta = \beta_0 + \beta_1 X + \beta_2 X^2 + \cdots + \beta_{p-1} X^{p-1}$.
- **Análisis categórico**. Para ello se categoriza la variable X considerando m nodos, es decir, si la covariable está definida en un intervalo $[a, b]$ se divide este en subintervalos de la forma:

  $$a = k_1 < k_2 < \cdots < k_{m-1} < k_m = b$$

  Se crean $m-2$ variables artificiales o dummy, $D_1, \ldots, D_{m-2}$, es decir, variables que toman el valor 0 ó 1 en función de la variable explicativa no se encuentre o sí en el i-ésimo subintervalo. Por último, se ajusta el modelo:

  $$\eta = \beta_0 + \beta_1 D_1 + \cdots + \beta_{m-2} D_{m-2}$$

  El análisis categórico presenta algunos inconvenientes, entre ellos: el ajuste final no es lineal, presenta un efecto constante en cada categoría y la selección de los nodos es arbitraría.

- **Regresión spline**. Esta regresión se introduce para resolver los problemas anteriores, pues realiza un ajuste suave y detecta los comportamientos locales.

  En ella se considera una base de $m-1$ polinomios a trozos de grado l, los cuales se ajustan por separado en cada uno de los intervalos definidos por una secuencia de nodos. A estos polinomios se le llaman bases. Para asegurar la continuidad de la estimación final de f, en las fronteras de los intervalos se definen las siguientes clases de bases de funciones.

  Definición 4.3.1. *Una función $f : [a, b] \to \mathbb{R}$ es un **spline polinómico** de grado l si satisface las siguientes condiciones:*

  - *$f(x)$ es $l-1$ veces continuamente diferenciable.*
  - *$f(x)$ es un polinomio de grado l para $x \in [k_j, k_{j+1}]$, con $j = 1 \ldots, m-1$.*

  Por lo tanto, cada spline polinómico puede ser ajustado por una base de $d = (m + 1 - l)$ funciones de la siguiente manera:

  $$f(x) = \sum_{j=1}^{d} \beta_j B_j(x)$$

  El spline polinómico que se utiliza con más frecuencia es el **cubic spline** o **spline cúbico**, $d = 3$. Entonces, se dirá que un spline polinómico, $f(x)$, es un cubic spline si:

  - $f(x)$ es 2 veces continuamente diferenciable.
  - $f(x)$ es un polinomio de grado 3 para $x \in [k_j, k_{j+1}]$, con $j = 1 \ldots, m-1$.

B-splines

Una posible base local es la formada por los Basic-splines, a los que llamaremos a partir de ahora B-splines. Los B-splines de grado l se obtienen al combinar $l+1$ polinomios de grado l en los $l-1$ nodos interiores. Por lo que un B-spline de grado 0, es decir, $l=0$, se define como:

$$B_j^0 = \mathbf{I}_{[k_j,k_{j+1})}(x) = \begin{cases} 1 & \text{si } x \in [k_j, k_{j+1}) \\ 0 & \text{en otro caso} \end{cases}$$

Los B-splines de grado mayor se definen recursivamente como:

$$B_j^l(x) = \frac{x-k_j}{k_{j+l}-k_j} B_j^{l-1}(x) + \frac{k_{j+l+1}-x}{k_{j+l+1}-k_{j+1}} B_{j+1}^{l-1}(x)$$

Supongamos ahora que queremos estimar el modelo de regresión:

$$\eta = f(x), \text{ siendo } f \text{ un función suave desconocida}$$

Para ello, se construirá una base de $m+l-1$ B-splines, $\{B_1(x), B_2(x), \ldots, B_{m+l-1}(x)\}$, donde m denota el número de nodos y l el grado de los polinomios, normalmente se considera $l=3$. Se evalúan los B-splines, B_j, en cada valor de la covariable X y se ajusta el siguiente modelo GLM:

$$\eta = \sum_{j=1}^{d} \beta_j B_j(x), \text{ donde } d = m+l-1$$

P-splines

Nótese que, en el caso anterior, tanto el valor de m como el de l son elegidos por el investigador. Además, cuantos más grados de libertad, más nodos, por lo que habrá mayor flexibilidad, es decir, las curvas serán demasiado "ruidosas", y viceversa, cuanto menos grados de libertad, menor número de nodos, por lo que será más suave el ajuste, es decir, puede que no capture la variabilidad de los datos. Por ese motivo, será necesario seleccionar un parámetro óptimo de suavización. Para resolver este problema, se introduce la regresión spline penalizada, a la que llamaremos a partir de ahora **P-spline**.

Para ello, se va a considerar un modelo que tenga una función de suavización y una covariable, es decir:

$$\eta = f(x) = \sum_{j=1}^{d} \beta_j B_j(x) \tag{4.1}$$

Para controlar la suavización se va a añadir un parámetro de penalización, λ, sobre el vector de coeficientes, para así, penalizar la gran variabilidad de f. A esto se le conoce como regresión spline penalizada.

Entonces, si se quiere estimar los parámetros del modelo [4.1], la regresión spline penalizada consiste en minimizar la deviance penalizada, que consiste en añadirle a la deviance una penalización sobre la curvatura de la función:

$$\| \sqrt{W}(Z - X\beta) \|^2 + \lambda \int \left(f''(x) \right)^2 dx$$

Al ser f una función lineal en los parámetro β, la penalización se puede expresar como una forma cuadrática, es decir:

$$\int \left(f^{''}(x)\right)^2 dx = \beta^t K \beta$$

donde K es la matriz de penalización, en la cual, los parámetros son conocidos.

De este modo, la función a minimizar será:

$$\| \sqrt{W}(Z - X\beta) \|^2 + \lambda \beta^t K \beta$$

Esta minimización se realiza utilizando el algoritmo de P-IRLS, el cual, fijando los parámetros de suavización es análogo al algoritmo de Fisher scoring introduciendo penalización en los parámetros. Por lo que dado un λ, $\beta^{(k+1)} = \left(X^t W^{(k)} X + \lambda K\right)^{-1} X^t W^{(k)} z^{(k)}$.

Para ver los pasos que sigue este algoritmo puede consultarse Wood (2006) [14].

Los estimadores P-IRLS tienen las siguientes propiedades:

- Dado un λ, el suavizador spline penalizado $\widehat{f}$ viene caracterizado por su matriz suavizadora:
$$S_\lambda = X \left(X^t W X + \lambda K\right)^{-1} X^t W$$
- Dado un λ, el suavizador spline penalizado $\widehat{f}$ es un suavizador "lineal": $\widehat{f} = S_\lambda z$.

En este punto surge un problema y es cómo seleccionar el parámetro adecuado de penalización óptimo, λ_{opt}. Para ello, se pueden utilizar los siguientes criterios automáticos de selección del grado de suavización:

- Generalized Cross-Validation (GCV). Se suele utilizar para modelos con parámetros escala desconocidos y tiene la siguiente forma:
$$GCV(\lambda) = \frac{n \times Deviance}{\left(n - \gamma tr(S_\lambda)\right)^2}, \text{ donde } \gamma \text{ es el parámetro escala para evitar } \lambda \text{ saltos.}$$
- Akaike Information Crtiterion (AIC).
- Umbiased Risk Estimator (UBRE). Se utiliza para modelos con parámetros de escala conocidos y tiene la siguiente forma:
$$UBRE(\lambda) = \frac{Deviance}{n} + \frac{2\gamma\phi tr(S_\lambda)}{n} - \phi$$
- Restricted Maximum Likelihood (REML). Cualquier suavizador penalizado cuadráticamente admite una representación como un modelo mixto de la forma:
$$X_F \beta_F + Zb \text{ donde } b \in N(0, I/\lambda)$$
Esta representación permite examinar el problema de selección del parámetro de suavización desde un perspectiva diferente.

  Para más información sobre estos criterios puede consultarse los apartados 6.2.5 y 6.2.6 de Wood (2006) [14].

En la práctica, y dependiendo de la base seleccionada, $\{B_j(x)\}$, existen varios suavizadores penalizados. Entre todos ellos, se han visto:

- Cubic regression spline.
- P-spline.

Existen otras como thin plate regression spline, adaptive smoothers,... Para más información sobre ellas se puede consultar Wood (2006) [14] y la página 246 de la documentación del paquete **mgcv** de **R** [5].

4.4. Modelo aditivo generalizado (GAM)

Al realizar una regresión spline se seleccionan los parámetros de una manera subjetiva, y además, puede ocurrir que los efectos tengan una forma desconocida y que las interacciones entre las covariables adopten una forma compleja. Por ese motivo se introducen los **modelos aditivos generalizados**, GAM.

Un modelo aditivo generalizado (GAM) es un modelo lineal generalizado (GLM), en el cual se reemplaza el predictor estrictamente lineal del GLM por un predictor aditivo semiparamétrico de la forma:

$$\eta = g(\mu) = \mathbf{X}^*\boldsymbol{\theta} + f_1(x_1) + f_2(x_2) + \cdots \tag{4.2}$$

donde $\mathbf{X}^*$ es la matriz de las componentes estrictamente paramétricas, α es el correspondiente vector de parámetros y $f_j(x_j)$ es el efecto parcial suave de x_j en el predictor.

Al igual que en la sección 4.3, cada función suavizadora, f_j, se puede representar como una combinación de una base de funciones $\{B_j\}$. De modo que f_j se puede reescribir como sigue:

$$f_j(X_j) = \sum_{k=1}^{q_j} \beta_{jk} B_{jk}(X_j)$$

La expresión matricial para el j-ésimo elemento sería:

$$f_j = \widetilde{X}_j \widetilde{B}_j$$

donde $\widetilde{X}_j = (B_{j,1}, \ldots, B_{j,q_1})$ y $\widetilde{B}_j = \left(\beta_{j,1}, \ldots, \beta_{j,q_j}\right)^t$.

Por lo tanto:

$$\begin{cases} \eta = g(\mu) = X^*\theta + \widetilde{X}_1\widetilde{B}_1 + \widetilde{X}_2\widetilde{B}_2 + \cdots \\ s.a. \quad 1^t\widetilde{X}_j\widetilde{B}_j = 0 \\ \qquad\quad 1^t\widetilde{X}_j Z = 0 \end{cases}$$

siendo Z una matriz con $q_j - q$ columnas ortogonales.

Es necesario introducir las restricciones, puesto que si no se hace el modelo no estaría identificado. La primera de las restricciones impone que la suma de los elementos de las funciones f_j es cero y la segunda que podemos encontrar una matriz Z con $q_j - 1$ columnas ortogonales que satisfaga esa condición.

Si reparametrizamos la función suave en términos de los $q_j - 1$ parámetros, β_j, tal que: $X_j = \widetilde{X}_j Z$, obtenemos una nueva matriz del modelo para el término j-ésimo, tal que:

$$f_j = X_j\beta_j$$

la cual satisface la restricción de estar centrada.

Dadas las matrices centradas por cada término suave, el modelo GAM se puede expresar como:

$$\eta = g(\mu) = X\beta \tag{4.3}$$

donde $X = (X^*, X_1, X_2, \ldots)^t$ y $\beta = (\theta, \beta_1, \beta_2, \ldots)$.

Supongamos que los parámetros de penalización son conocidos y que queremos estimar los parámetros del modelo (4.3). Al igual que en la sección 4,3, la estimación del parámetro β se lleva a cabo minimizando la deviance penalizada, es decir, minimizando:

$$\| \sqrt{W}(z - X\beta) \|^2 + \lambda\beta^t K\beta = \| \sqrt{W}(z - X\beta) \|^2 + \sum_{j=1}^{p} \lambda_j \beta_j^t K_j \beta_j$$

Del mismo modo, la minimización se realiza utilizando el algoritmo de P-IRLS. Obteniendo así, que el estimador del parámetro β es:

$$\widehat{\beta}\,(X^t W X + s)^{-1} X^t W z,$$

donde $S = \sum_{j=1}^{p} \lambda_j K_j$.

Tanto los criterios de selección del parámetro de penalización óptimo como las bases de suavización a utilizar son las mismas que en la sección 4.3.

Nótese que tanto los modelo GLM de la sección 4.3 como los modelo GAM son válidos si los residuos de dichos modelos pasan los contrastes de normalidad, homocedasticidad e independencia. Si el modelo es válido se podrá realizar predicciones a futuro. En caso de que no lo sea se deberá ajustar otro modelo hasta que este sea válido.

En este caso, al estar utilizando series temporales, es muy posible que los errores estén correlados. En este caso se propondrá un modelo ARMA para ellos y se ajustarán conjuntamente los parámetros de la regresión y del ARMA. En la práctica, se llevará a cabo esta idea utilizando la función *gamm* de la librería *mgcv* de *R* [5]. Veamos un ejemplo para saber cómo utilizar esta función. Para ello, se considerarán como variables una única variable explicativa y una variable respuesta, las cuales se muestran en la figura 4.2.

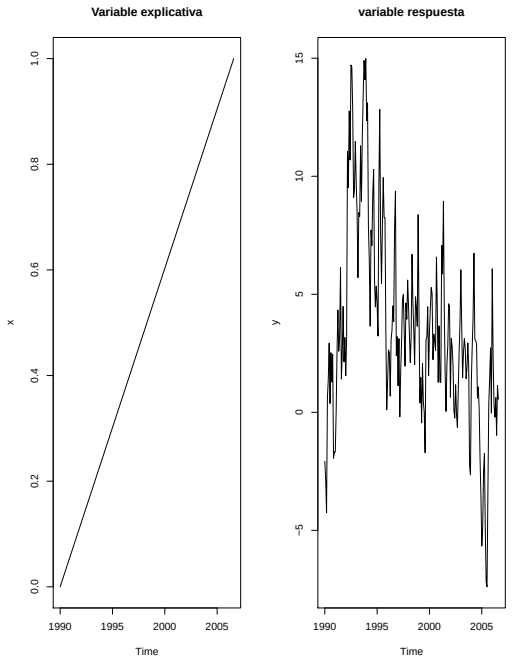


Figura 4.2: Series temporales: variable explicativa (izquierda) y variable respuesta (derecha)

Se ajusta un GAM, utilizando como base suavizadora la cubic regression spline y como método de estimación de los parámetros el REML. Una vez ajustado el modelo se verá si este es válido, realizando para ello los contrastes mencionados anteriormente. Observando el cuadro 4.1 y considerando como nivel de significación $\alpha = 0.05$, se llega a la conclusión de que el modelo no es válido, pues los residuos de dicho modelo no pasan el contraste de independencia, por lo que están correlados ya que bajo normalidad la independencia equivale a incorrelación.

Hipótesis	p-valor
Normalidad	0,4245
Independencia	7,883 10^{-15}

Cuadro 4.1: Contrastes sobre los errores del modelo

Por lo tanto, se deberá ajustar un ARMA a los residuos del modelo. Observando el gráfico de la fas y fap de los residuos del modelo GAM, figura 4.3, se puede sugerir que los residuos siguen un modelo $AR(1)$.

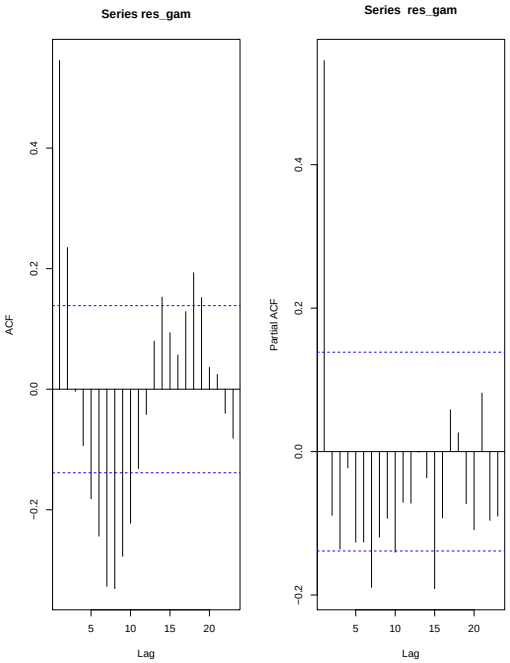


Figura 4.3: fas y fap de los residuos del modelo

Nótese que no se ha comprobado si el modelo propuesto para los residuos es válido, pues lo que interesa es que el modelo final sí lo sea. Entonces si el modelo final, entendiendo como modelo final aquel en el que se ajusta conjuntamente los parámetros de la regresión y el $ARMA$, supera los contrastes sobre los residuos, se habrá encontrado un modelo válido. Para la construcción del modelo final utilizaremos la función *gamm*, mencionada anteriormente, indicándole que los residuos siguen un $AR(1)$, es decir, nuestro modelo sería:

$$mod = gamm(y \sim s(x, bs = \text{“}cr\text{”}), method = \text{“}REML\text{”}, correlation = corARMA(p = 1, q = 0))$$

Los contrastes de los residuos del modelo final se pueden observar en el cuadro 4.2. Si se toma como nivel de significación $\alpha = 0{,}05$ se observa que los residuos pasan tanto el contraste de normalidad como

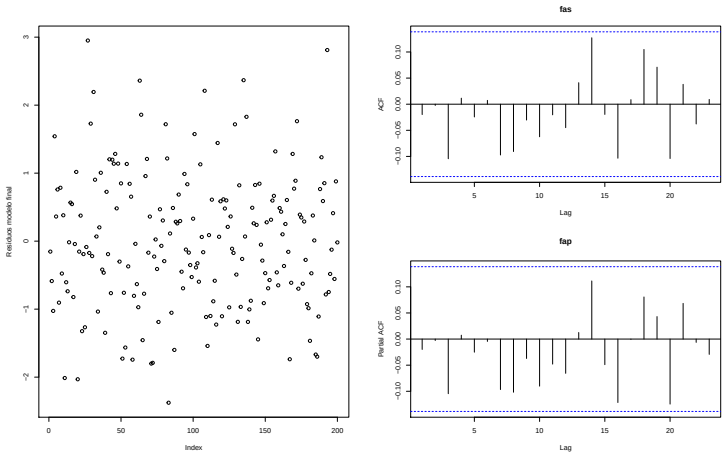
en el de independencia, pues en ambos casos el p-valor es mayor que el nivel de significación, por lo que no hay evidencias suficientes para rechazar la hipótesis nula. El contraste de estacionariedad se realiza mediante la función *adf.test* del paquete *aTSA* de *R*. En dicho contraste la hipótesis nula es que los datos no son estacionarios frente a la hipótesis alternativa de que sí lo son. Como el p-valor obtenido es menor que el nivel de significación, rechazaremos la hipótesis nula, siendo así, los residuos del modelo final estacionarios.

Hipótesis	p-valor
Normalidad	0,318
Independencia	0,7772
Estacionariedad	0,01

Cuadro 4.2: Contrastes sobre los errores del modelo final

La gráfica de los residuos del modelo final, figura 4.4a, muestra que estos son homocedásticos y las gráficas de las fas y fap de dichos residuos, figura 4.4b, nos indican que son ruido blanco.

Por lo tanto, podemos concluir, con todo lo dicho anteriormente, que el modelo final es válido. De este modo, si se quiere, se podrán hacer predicciones a futuro.



(a) Residuos del modelo final

(b) fas y fap de los residuos del modelo final

Figura 4.4: Residuos del modelo final

Parte II

Caso práctico

Capítulo 5

Preparación de los datos para el análisis

En esta parte del trabajo se utilizarán las series temporales del PIB, del consumo de hogares gallegos y las series internas de ABANCA, las cuales han sido descritas en el capítulo 1. Como se ha mencionado en dicho capítulo, la frecuencia de las series del PIB y del consumo de hogares gallegos es trimestral, mientras que la frecuencia de las series internas es mensual.

Para poder llevar a cavo un análisis es necesario que la frecuencia de observación de todas las variables implicadas en el análisis sea la misma. Por ese motivo se procederá al cambio de frecuencia de las series del PIB real, en volumen, con base 2010 y del consumo de hogares gallegos, es decir, pasarán de tener una frecuencia trimestral a una frecuencia mensual. Esta desagregación se realizará utilizando la función:

$$mod \leftarrow td(pib \sim 1, to = \text{“}monthly\text{”}, method = \text{“}denton - cholette\text{”}, conversion = \text{“}average\text{”})$$

de la librería *tempdisagg*. Dicha función realiza una regresión donde la variable respuesta es la serie de baja frecuencia y las variables explicativas los indicadores. En este caso no se tienen indicadores, de modo que la variable explicativa es 1. Para más información sobre dicha función puede consultarse [6].

Al realizar la predicción del modelo, es decir, al ejecutar la sentencia $pibm \leftarrow predict(mod, length(pib) * 3)$, se obtiene la desagregación de la serie, en este caso, la serie del PIB con frecuencia mensual, sujeto a que la media de un trimestre de la serie desagregada coincida con el valor del trimestre de la serie de baja frecuencia, es decir, dada la serie trimestral del PIB real, en volumen, con base 2010:

	Qtr1	Qtr2	Qtr3	Qtr4
2007			102.62	103.39
2008	103.94	104.19	103.74	102.49
2009	100.89	99.55	99.29	99.51
2010	99.78	100.20	100.13	99.89
2011	99.64	99.20	98.36	97.65
2012	96.72	95.68	95.22	94.64
2013	94.74	95.05	95.21	95.67
2014	95.95	96.07	96.68	97.08
2015	97.67	98.51	99.08	99.95
2016	100.67	101.42	102.32	103.00
2017	103.66	104.62	105.45	106.15
2018	106.84			

y su desagregación:

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
2007							102.33274	102.62514	102.90212	103.16367	103.39881	103.60752
2008	103.78980	103.94802	104.08217	104.19226	104.21809	104.15966	104.01697	103.77357	103.42946	102.98464	102.50246	101.98290
2009	101.42598	100.88477	100.35925	99.84944	99.49741	99.30316	99.26669	99.27505	99.32826	99.42630	99.51358	99.59012
2010	99.65590	99.76542	99.91868	100.11567	100.22817	100.25617	100.19967	100.13329	100.05703	99.97088	99.88868	99.81043
2011	99.73613	99.64546	99.53842	99.41501	99.22290	98.96209	98.63257	98.34576	98.10167	97.90030	97.66223	97.38746
2012	97.07600	96.73113	96.35287	95.94121	95.64239	95.45640	95.38323	95.24252	95.03425	94.75843	94.60065	94.56092
2013	94.63924	94.73439	94.84637	94.97517	95.06349	95.11134	95.11871	95.18902	95.32227	95.51845	95.68116	95.81039

```
2014  95.90614  95.96297  95.98088  95.95986  96.03721  96.21292  96.48700  96.70027  96.85273  96.94438  97.06901  97.22661
2015  97.41719  97.65444  97.93837  98.26896  98.53239  98.72865  98.85773  99.05670  99.32557  99.66432  99.96327 100.22242
2016 100.44177 100.66778 100.90045 101.13977 101.40977 101.71045 102.04181 102.33329 102.58489 102.79661 103.00208 103.20130
2017 103.39427 103.64181 103.94392 104.30059 104.62932 104.93010 105.20293 105.45644 105.69062 105.90548 106.14258 106.40194
2018 106.68355 106.87129 106.96516
2018
```

La media, por ejemplo, de los primeros tres meses de 2008 tendría que coincidir con el PIB del primer trimestre de 2008. Esto es así, pues:

$$\frac{103{,}78980 + 103{,}94802 + 104{,}08217}{3} = 103{,}94$$

coincide con el PIB del primer trimestre del año 2008.

En la figura 5.1 se puede observar la serie trimestral del PIB, desde el tercer trimestre de 2007 en adelante, y la serie desagregada, es decir, la serie mensual desde Julio de 2007 a Marzo de 2018.

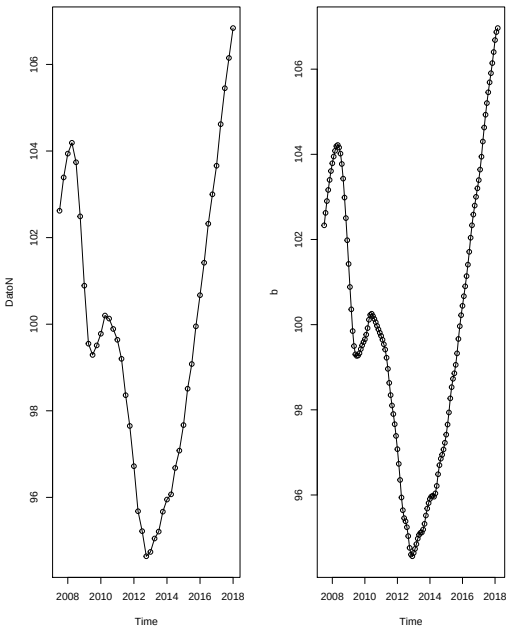


Figura 5.1: Gráfica del PIB (izquierda) y gráfica del PIB desagregado (derecha)

De manera análoga se desagregaría la serie referente al consumo de hogares gallegos. Nótese que ambas series, tanto la del PIB como la del consumo de hogares gallegos, comienzan en el tercer trimestre de 2007. Esto es así pues, se disponen de datos internos de ABANCA a partir de Julio de 2007.

Por otro lado, como se puede observar en las figuras, **eliminadas por confidencialidad**, casi todas las variables internas de ABANCA presentan un salto muy pronunciado entre Noviembre de 2010 y Diciembre del mismo año, debido a la fusión de las entidades bancarias Caixa Galicia y Caixanova. Por ese motivo, se van a corregir dichas series, calculando para ello el salto, es decir, la diferencia entre el valor en Diciembre de 2010 de la serie y el valor en Noviembre de 2010 de la misma series, y dicho valor se suma al "primer trozo de la serie" para levantarlo, pues los últimos años de la serie son los de mayor interés. En otras palabras, se calcula una constante α como $x_t - x_{t-1}$, siendo t Diciembre de 2010 y x la variable a corregir. Entonces la serie corregida en el instante t, x_{corr} en t, será:

$$x_{corr}\text{en t } = \begin{cases} x_t - \alpha & \text{si } x < \text{Diciembre 2010} \\ 0 & \text{en otro caso} \end{cases}$$

El proceso descrito se puede observar en la figura, **eliminada por confidencialidad**, donde la figura, **eliminada por confidencialidad**, muestra un serie con un salto entre Noviembre de 2010 y Diciembre de 2010 y la figura, **eliminada por confidencialidad**, muestra la serie corregida.

En las figuras, **eliminadas por confidencialidad**, se pueden observar las series corregidas. Nótese que en algunas se han corregido más de un salto, usando para ello, un proceso análogo al descrito anteriormente.

Figuras eliminadas por confidencialidad.

Figuras eliminadas por confidencialidad.

Figuras eliminadas por confidencialidad.

Figuras eliminadas por confidencialidad.

Capítulo 6

Obtención de las componentes principales dinámicas

En este capítulo se va a explicar el procedimiento utilizado para la obtención de las componentes principales dinámicas. Para ello se va a suponer que se está manejando un conjunto de datos internos, el cuál es una matriz, que llamaremos $matriz2\_correxida\_est$, donde cada columna representa una serie. Como se ha dicho anteriormente, se están a considerar trece variables internas. Por ello será de gran interés poder reducir la dimensión del conjunto de series temporales.

Al estar trabajando con series temporales no es adecuado utilizar el análisis de componentes principales para reducir la dimensión, sino, como se ha visto anteriormente, se debe utilizar el análisis de componentes principales dinámicas.

Para la obtención de las componentes principales dinámicas se ha seguido el artículo de Peña y Yohay [12]. Estos autores, junto a Ezequiel Smucler, han implementado una función en R con la que se obtienen las componentes principales dinámicas. Dicha función se encuentra en el paquete *gdpc* [4] y es:

```
fit<-auto.gdpc(Z=matriz2_correxida_est, crit = 'LOO', k_max =3, niter_max = 1500,
               normalize = 1,expl_var=0.999, tol = 1e-4)
```

siendo:

- Z: matriz en la que cada columna es una serie temporal. En nuestro caso las series se han corregido y estandarizado.
- crit: criterio para elegir el número adecuado de retardos.
- k_max: número máximo de los posibles retardos. Por defecto es 10 pero si $k\_max = 3$ se obtienen las compoenetes principales en un menor tiempo.
- niter_max: número máximo de iteraciones
- $normalized = 1$: indica que se quiere trabajar con las unidades originales de Z.
- expl_var: indica la proporción de varianza explicada.
- tol: indica la precisión.

Una vez calculadas las componentes principales, sino se tiene un valor para cada instante temporal del año 2018, será necesario predecirlo. Pues lo que se desea es obtener una predicción, tanto del PIB como del consumo de hogares gallegos, para el año 2018. Para obtener dichas predicciones se utilizará la metodología Box-Jenkins explicada en el capítulo 3, es decir, se ajustará un modelo AR, MA, $ARMA$ o $ARIMA$, se elegirá aquel que sea válido y, con el modelo seleccionado se realizarán predicciones para instantes futuros.

Enero

Supongamos que nos encontramos a principios de año, es decir, en el mes de Enero. En este instante se dispone de datos para cada una de las variables internas desde Julio del 2007 a Diciembre del 2017. Al estar considerando trece variables se procederá a la reducción de la dimensión, utilizando para ello el análisis de componentes principales dinámicas. En este caso obtenemos la siguiente información sobre las componentes principales dinámicas:

```
            Number.of.lags   LOO   MSE Explained.Variance
Component 1              3 0.133 0.122              0.878
Component 2              2 0.049 0.046              0.954
Component 3              3 0.021 0.019              0.981
Component 4              2 0.011 0.010              0.990
Component 5              3 0.006 0.006              0.994
Component 6              3 0.004 0.004              0.996
Component 7              3 0.002 0.002              0.998
Component 8              3 0.002 0.001              0.999
Component 9              3 0.001 0.001              0.999
```

Siguiendo lo expuesto en el apartado 2.1.3, llegaría con considerar las dos primeras componentes para reducir la dimensión del conjunto de variables explicativas, pues estas explican más del 90 % de la variabilidad. En la figura 6.1 se pueden observar las componentes principales.

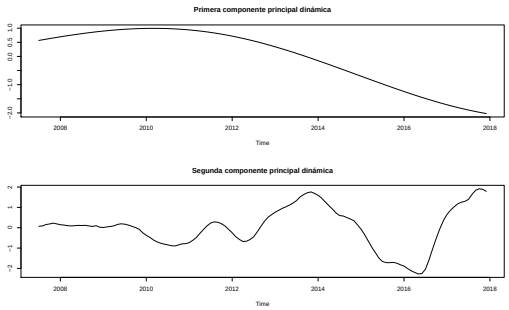


Figura 6.1: Primera componente principal dinámica (gráfica superior) y segunda componente principal dinámica (gráfica inferior) en Enero

Como se necesitan los valores de ambas componentes principales dinámicas en el año 2018 para la predicción del consumo de hogares gallegos y del PIB, será necesario ajustar un modelo, utilizando la metodología Box-Jenkins, para poder predecir dichos valores. El cuadro 6.1 muestra la validación del modelo utilizado para la predicción de los valores futuros. Aunque el modelo no verifique la hipótesis de normalidad este es válido pero no se podrán calcular los intervalos de confianza para los valores predichos. Las predicciones de las componentes para el año 2018 se pueden observar en la figura 6.2.

	c1	c2
Independencia	1	1
Media cero	0,8647	0,6864
Normalidad	$2{,}2e-16$	0,02322
Estacionariedad	0,01	0,01

Cuadro 6.1: Validación de los modelos ajustados para la primera componente principal dinámica, c1, y para la segunda componente principal dinámica, c2. Se incluyen los p-valores de los contrastes sobre las hipótesis en las que se basa el modelo.

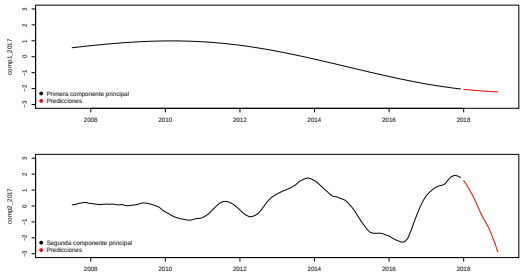


Figura 6.2: Primera componente principal dinámica, en negro, con sus predicciones para el 2018, en rojo, (gráfica superior) y segunda componente principal dinámica, en negro, con sus predicciones para el 2018, en rojo, (gráfica inferior), en Enero

Febrero

Supongamos que nos encontramos en el mes de Febrero. En este instante se dispone de datos para cada una de las variables internas desde Julio del 2007 a Enero del 2018. Al estar considerando trece variables se procederá a la reducción de la dimensión, utilizando para ello el análisis de componentes principales dinámicas. En este caso obtenemos la siguiente información sobre las componentes principales dinámicas:

```
             Number.of.lags   LOO   MSE Explained.Variance
Component 1               3 0.131 0.120              0.880
Component 2               3 0.040 0.037              0.963
Component 3               3 0.021 0.019              0.981
Component 4               3 0.011 0.010              0.990
Component 5               3 0.006 0.006              0.994
Component 6               3 0.004 0.003              0.997
Component 7               3 0.002 0.002              0.998
Component 8               3 0.002 0.001              0.999
Component 9               3 0.001 0.001              0.999
```

Siguiendo lo expuesto en el apartado 2.1.3, llegaría con tomar las dos primeras componentes para reducir la dimensión del conjunto de variables explicativas, pues estas, explican más del 90 % de la variabilidad. En la figura 6.3 se pueden observar las componentes principales.

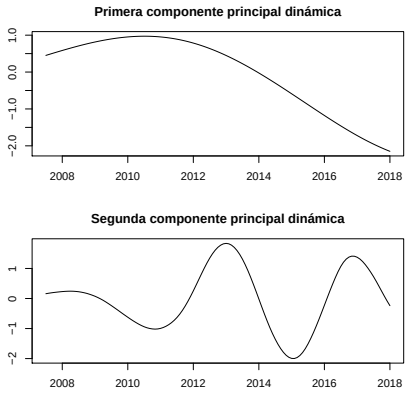


Figura 6.3: Primera componente principal dinámica (gráfica superior) y segunda componente principal dinámica (gráfica inferior) en Febrero

Para la predicción del consumo de hogares gallegos y del PIB en el año 2018 será necesario predecir los valores, en este caso, de las componentes principales dinámicas para los instantes temporales comprendidos entre Febrero y Diciembre de dicho año. Para ello se ajustará un modelo, utilizando la metodología Box-Jenkins. De este modelo se obtendrá una predicción de dichos valores. El cuadro 6.2 muestra la validación del modelo utilizado para la predicción de los valores futuros. Aunque el modelo no verifique la hipótesis de normalidad este es válido pero no se podrán calcular los intervalos de confianza para los valores predichos. Las predicciones de las componentes para el año 2018 se pueden observar en la figura 6.4.

	c1	c2
Independencia	1	1
Media cero	0,8953	0,9465
Normalidad	$2{,}2e-16$	0,4296
Estacionariedad	0,01	0,01

Cuadro 6.2: Validación de os modelos ajustados para la primera componente principal dinámica, c1, y para la segunda componente principal dinámica, c2. Se incluyen los p-valores de los contrastes sobre las hipótesis en las que se basa el modelo.

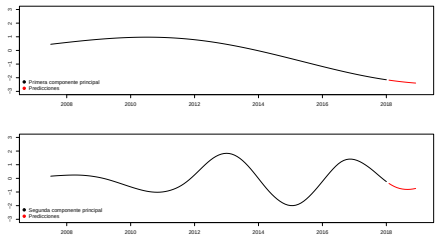


Figura 6.4: Primera componente principal dinámica, en negro, con sus predicciones para el 2018, en rojo, (gráfica superior) y segunda componente principal dinámica, en negro, con sus predicciones para el 2018, en rojo, (gráfica inferior), en Febrero

Marzo

Supongamos que nos encontramos en el mes de Marzo. En este instante se dispone de datos para cada una de las variables internas desde Julio del 2007 a Febrero del 2018. Al estar considerando trece variables se procederá a la reducción de la dimensión, utilizando para ello el análisis de componentes principales dinámicas. En este caso obtenemos la siguiente información sobre las componentes principales dinámicas:

```
        Number.of.lags   LOO   MSE Explained.Variance
Component 1              3 0.129 0.118              0.882
Component 2              3 0.040 0.037              0.963
Component 3              3 0.020 0.019              0.981
Component 4              3 0.011 0.010              0.990
Component 5              3 0.007 0.006              0.994
Component 6              2 0.004 0.004              0.996
Component 7              3 0.003 0.002              0.998
Component 8              2 0.002 0.002              0.998
Component 9              3 0.001 0.001              0.999
```

Siguiendo lo expuesto en el apartado 2.1.3, llegaría con considerar las dos primeras componentes para reducir la dimensión del conjunto de variables explicativas, pues estas explican más del 90 % de la variabilidad. En la figura 6.5 se pueden observar las componentes principales.

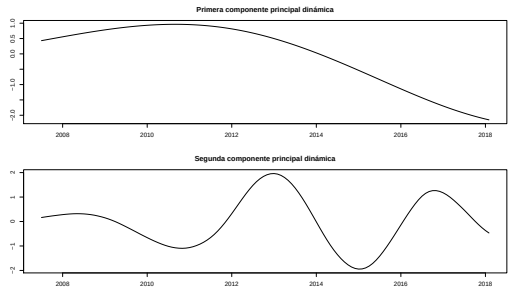


Figura 6.5: Primera componente principal dinámica (gráfico superior) y segunda componente principal dinámica (gráfico inferior) en Marzo

Para la predicción del consumo de hogares gallegos y del PIB en el año 2018 será necesario predecir los valores, en este caso, de las componentes principales dinámicas para los instantes temporales comprendidos entre Marzo y Diciembre de dicho año. Para ello se ajustará un modelo, utilizando la metodología Box-Jenkins. De este modelo se obtendrá una predicción de dichos valores. El cuadro 6.3 muestra la validación del modelo utilizado para la predicción de los valores futuros. Aunque el modelo no verifique la hipótesis de normalidad este es válido pero no se podrán calcular los intervalos de confianza para los valores predichos. Las predicciones de las componentes para el año 2018 se pueden observar en la figura 6.6.

	c1	c2
Independencia	1	1
Media cero	0,4626	0,8827
Normalidad	$2{,}2e-16$	0,21
Estacionariedad	0,01	0,01

Cuadro 6.3: Validación de os modelos ajustados para la primera componente principal dinámica, c1, y para la segunda componente principal dinámica, c2. Se incluyen los p-valores de los contrastes sobre las hipótesis en las que se basa el modelo.

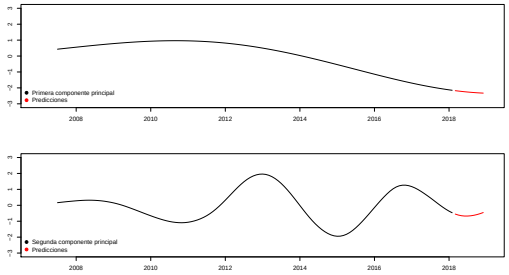


Figura 6.6: Primera componente principal dinámica, en negro, con sus predicciones para el 2018, en rojo, (gráfica superior) y segunda componente principal dinámica, en negro, con sus predicciones para el 2018, en rojo, (gráfica inferior), en Marzo

Abril

Supongamos que nos encontramos en el mes de Abril. En este instante se dispone de datos para cada una de las variables internas desde Julio del 2007 a Marzo del 2018. Al estar considerando trece variables se procederá a la reducción de la dimensión, utilizando para ello el análisis de componentes principales dinámicas. En este caso obtenemos la siguiente información sobre las componentes principales dinámicas:

```
     Number.of.lags   LOO   MSE Explained.Variance
Component 1              3 0.126 0.116             0.884
Component 2              3 0.039 0.035             0.965
Component 3              3 0.020 0.019             0.981
Component 4              3 0.010 0.010             0.990
Component 5              3 0.006 0.005             0.995
Component 6              3 0.004 0.003             0.997
Component 7              3 0.002 0.002             0.998
Component 8              3 0.001 0.001             0.999
Component 9              3 0.001 0.001             0.999
```

Siguiendo lo expuesto en el apartado 2.1.3, llegaría con seleccionar las dos primeras componentes para reducir la dimensión del conjunto de variables explicativas, pues estas explican más del 90 % de la variabilidad. En la figura 6.7 se pueden observar las componentes principales.

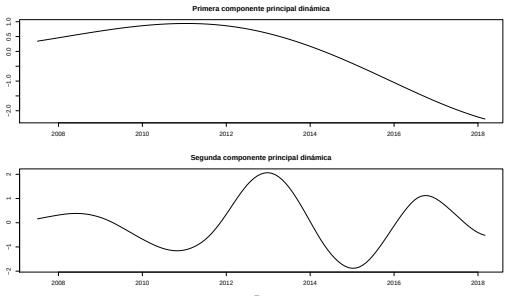


Figura 6.7: Primera componente principal dinámica (gráfico superior) y segunda componente principal dinámica (gráfico inferior) en Abril

Para la predicción del consumo de hogares gallegos y del PIB en el año 2018 será necesarios predecir los valores, en este caso, de las componentes principales dinámicas para los instantes temporales comprendidos entre Abril y Diciembre de dicho año. Para ello se ajustará un modelo, utilizando la metodología Box-Jenkins. De este modelo se obtendrá una predicción dichos valores. El cuadro 6.4 muestra la validación del modelo utilizado para la predicción de los valores futuros. Aunque el modelo no verifique la hipótesis de normalidad este es válido pero no se podrán calcular los intervalos de confianza para los valores predichos. Las predicciones de las componentes para el año 2018 se pueden observar en la figura 6.8.

	c1	c2
Independencia	1	1
Media cero	0,07763	0,8425
Normalidad	$2{,}2e-16$	0,2972
Estacionariedad	0,03981349	0,01

Cuadro 6.4: Validación de os modelos ajustados para la primera componente principal dinámica, c1, y para la segunda componente principal dinámica, c2. Se incluyen los p-valores de los contrastes sobre las hipótesis en las que se basa el modelo.

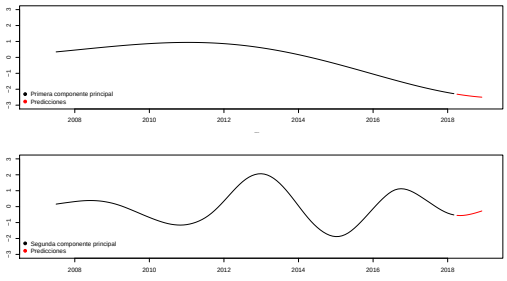


Figura 6.8: Primera componente principal dinámica, en negro, con sus predicciones para el 2018, en rojo, (gráfica superior) y segunda componente principal dinámica, en negro, con sus predicciones para el 2018, en rojo, (gráfica inferior), en Abril

Mayo

Supongamos que nos encontramos en el mes de Mayo. En este instante se dispone de datos para cada una de las variables internas desde Julio del 2007 a Abril del 2018. Al estar considerando trece variables se procederá a la reducción de la dimensión, utilizando para ello el análisis de componentes principales dinámicas. En este caso obtenemos la siguiente información sobre las componentes principales dinámicas:

	Number.of.lags	LOO	MSE	Explained.Variance
Component 1	3	0.123	0.114	0.886
Component 2	3	0.037	0.034	0.966
Component 3	2	0.020	0.019	0.981
Component 4	3	0.011	0.010	0.990
Component 5	3	0.006	0.005	0.995
Component 6	3	0.004	0.003	0.997
Component 7	3	0.002	0.002	0.998
Component 8	3	0.001	0.001	0.999
Component 9	3	0.001	0.001	0.999

Siguiendo lo expuesto en el apartado 2.1.3, llegaría con seleccionar las dos primeras componentes para reducir la dimensión del conjunto de variables explicativas, pues estas explican más del 90 % de la variabilidad. En la figura 6.9 se pueden observar las componentes principales.

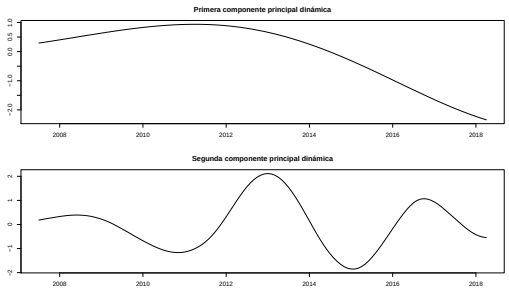


Figura 6.9: Primera componente principal dinámica (gráfico superior) y segunda componente principal dinámica (gráfico inferior) en Mayo

Para la predicción del consumo de hogares gallegos y del PIB en el año 2018 serán necesarios predecir los valores, en este caso, de las componentes principales dinámicas para los instantes temporales comprendidos entre Mayo y Diciembre de dicho año. Para ello se ajustará un modelo, utilizando la metodología Box-Jenkins. De este modelo se obtendrá una predicción dichos valores. El cuadro 6.5 muestra la validación del modelo utilizado para la predicción de los valores futuros. Aunque el modelo no verifique la hipótesis de normalidad este es válido pero no se podrán calcular los intervalos de confianza para los valore predichos. Las predicciones de las componentes para el año 2018 se pueden observar en la figura 6.10.

	c1	c2
Independencia	1	1
Media cero	0,1	0,9
Normalidad	$2{,}2e-16$	0,2676
Estacionariedad	0,02040611	0,01

Cuadro 6.5: Validación de os modelos ajustados para la primera componente principal dinámica, c1, y para la segunda componente principal dinámica, c2. Se incluyen los p-valores de los contrastes sobre las hipótesis en las que se basa el modelo.

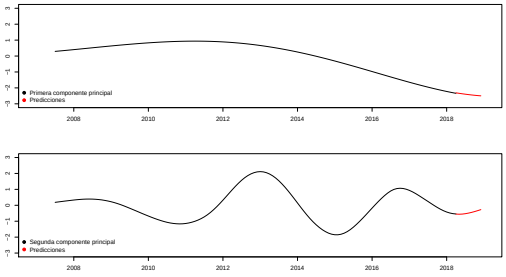


Figura 6.10: Primera componente principal dinámica, en negro, con sus predicciones para el 2018, en rojo, (gráfica superior) y segunda componente principal dinámica, en negro, con sus predicciones para el 2018, en rojo, (gráfica inferior), en Mayo

Junio

Supongamos que nos encontramos en el mes de Junio. En este instante se dispone de datos para cada una de las variables internas desde Julio del 2007 a Mayo del 2018. Al estar considerando trece variables se procederá a la reducción de la dimensión, utilizando para ello el análisis de componentes principales dinámicas. En este caso obtenemos la siguiente información sobre las componentes principales dinámicas:

```
             Number.of.lags   LOO   MSE Explained.Variance
Component 1               3 0.120 0.111              0.889
Component 2               2 0.045 0.042              0.958
Component 3               3 0.019 0.017              0.983
Component 4               3 0.009 0.008              0.992
Component 5               3 0.005 0.005              0.995
Component 6               3 0.003 0.003              0.997
Component 7               3 0.002 0.002              0.998
Component 8               3 0.001 0.001              0.999
Component 9               3 0.001 0.001              0.999
```

Siguiendo lo expuesto en el apartado 2.1.3, llegaría con considerar las dos primeras componentes para reducir la dimensión del conjunto de variables explicativas, pues estas explican más del 90 % de la variabilidad. En la figura 6.11 se pueden observar las componentes principales.

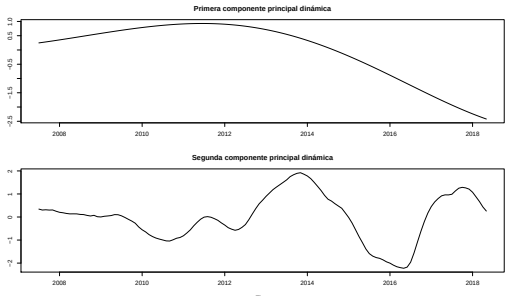


Figura 6.11: Primera componente principal dinámica (gráfico superior) y segunda componente principal dinámica (gráfico inferior) en Junio

Para la predicción del consumo de hogares gallegos y del PIB en el año 2018 será necesario predecir los valores, en este caso, de las componentes principales dinámicas para los instantes temporales comprendidos entre Junio y Diciembre de dicho año. Para ello se ajustará un modelo, utilizando la metodología Box-Jenkins. De este modelo se obtendrá una predicción de dichos valores. El cuadro 6.6 muestra la validación del modelo utilizado para la predicción de los valores futuros. Aunque el modelo no verifique la hipótesis de normalidad este es válido pero no se podrán calcular los intervalos de confianza para los valores predichos. Las predicciones de las componentes para el año 2018 se pueden observar en la figura 6.12.

	c1	c2
Independencia	1	1
Media cero	0,09599	0,9999
Normalidad	$2{,}2e-16$	0,4784
Estacionariedad	0,03499217	0,01

Cuadro 6.6: Validación de os modelos ajustados para la primera componente principal dinámica, c1, y para la segunda componente principal dinámica, c2. Se incluyen los p-valores de los contrastes sobre las hipótesis en las que se basa el modelo.

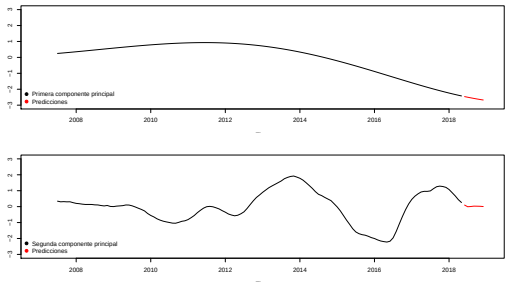


Figura 6.12: Primera componente principal dinámica, en negro, con sus predicciones para el 2018, en rojo, (gráfica superior) y segunda componente principal dinámica, en negro, con sus predicciones para el 2018, en rojo, (gráfica inferior), en Junio

Julio

Supongamos que nos encontramos en el mes de Julio. En este instante se dispone de datos para cada una de las variables internas desde Julio del 2007 a Junio del 2018. Al estar considerando trece variables se procederá a la reducción de la dimensión, utilizando para ello el análisis de componentes principales dinámicas. En este caso obtenemos la siguiente información sobre las componentes principales dinámicas:

	Number.of.lags	LOO	MSE	Explained.Variance
Component 1	3	0.106	0.098	0.902
Component 2	3	0.045	0.041	0.959
Component 3	3	0.021	0.020	0.980
Component 4	2	0.012	0.012	0.988
Component 5	3	0.007	0.007	0.993
Component 6	2	0.005	0.004	0.996
Component 7	3	0.003	0.003	0.997
Component 8	3	0.002	0.002	0.998
Component 9	3	0.001	0.001	0.999
Component 10	2	0.001	0.001	0.999

Siguiendo lo expuesto en el apartado 2.1.3, llegaría con seleccionar la primera componente para reducir la dimensión del conjunto de variables explicativas, pues esta explican más del 90 % de la variabilidad. Sin embargo, para el caso práctico se considerarán las dos primeras componentes principales. En la figura 6.13 se pueden observar las componentes principales.

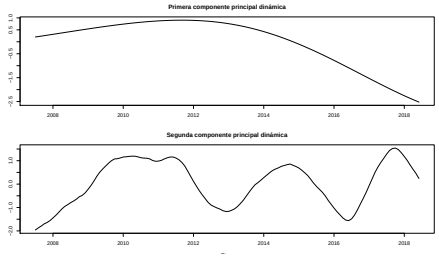


Figura 6.13: Primera componente principal dinámica (gráfico superior) y segunda componente principal dinámica (gráfico inferior) en Julio

Para la predicción del consumo de hogares gallegos y del PIB en el año 2018 será necesario predecir los valores, en este caso, de las componentes principales dinámicas para los instantes temporales comprendidos entre Julio y Diciembre de dicho año. Para ello se ajustará un modelo, utilizando la metodología Box-Jenkins. De este modelo se obtendrá una predicción de dichos valores. El cuadro 6.7 muestra la validación del modelo utilizado para la predicción de los valores futuros. Aunque el modelo no verifique la hipótesis de normalidad este es válido pero no se podrán calcular los intervalos de confianza para los valores predichos. Las predicciones de las componentes para el año 2018 se pueden observar en la figura 6.14.

	c1	c2
Independencia	1	1
Media cero	0,7333	0,8166
Normalidad	$2{,}2e-16$	0,9666
Estaconariedad	0,02670112	0,01639

Cuadro 6.7: Validación de os modelos ajustados para la primera componente principal dinámica, c1, y para la segunda componente principal dinámica, c2. Se incluyen los p-valores de los contrastes sobre las hipótesis en las que se basa el modelo.

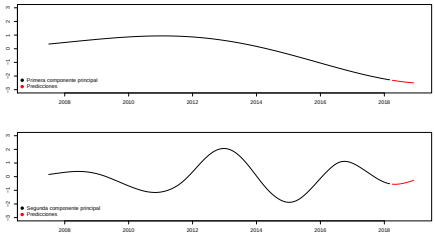


Figura 6.14: Primera componente principal dinámica, en negro, con sus predicciones para el 2018, en rojo, (gráfica superior) y segunda componente principal dinámica, en negro, con sus predicciones para el 2018, en rojo, (gráfica inferior), en Julio

Agosto

Supongamos que nos encontramos en el mes de Agosto. En este instante se dispone de datos para cada una de las variables internas desde Julio del 2007 a Julio del 2018. Al estar considerando trece variables se procederá a la reducción de la dimensión, utilizando para ello, el análisis de componentes principales dinámicas. En este caso, obtenemos la siguiente información sobre las componentes principales dinámicas:

```
        Number.of.lags   LOO   MSE Explained.Variance
Component 1                 3 0.100 0.092              0.908
Component 2                 3 0.039 0.036              0.964
Component 3                 3 0.024 0.022              0.978
Component 4                 3 0.012 0.011              0.989
Component 5                 3 0.008 0.007              0.993
Component 6                 3 0.005 0.004              0.996
Component 7                 3 0.003 0.003              0.997
Component 8                 3 0.002 0.002              0.998
Component 9                 3 0.001 0.001              0.999
Component 10                3 0.001 0.001              0.999
```

Siguiendo lo expuesto en el apartado 2.1.3, llegaría con considerar la primera componente para reducir la dimensión del conjunto de variables explicativas, pues esta explica más del 90 % de la variabilidad. Sin embargo, para el caso practico se seleccionarán las dos primeras componentes. En la figura 6.15 se puede observar las componentes principales.

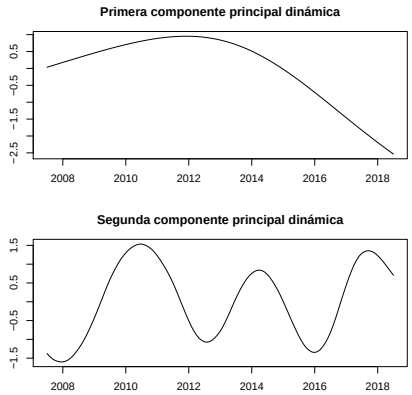


Figura 6.15: Primera componente principal dinámica (gráfico superior) y segunda componente principal dinámica (gráfico inferior) en Agosto

Para la predicción del consumo de hogares gallegos y del PIB en el año 2018 será necesario predecir los valores, en este caso de las componentes principales dinámicas para los instantes temporales comprendidos entre Agosto y Diciembre de dicho año. Para ello se ajustará un modelo, utilizando la metodología Box-Jenkins. De este modelo se obtendrá una predicción de dichos valores. El cuadro 6.8 muestra la validación del modelo utilizado para la predicción de los valores futuros. Aunque el modelo no verifique la hipótesis de normalidad este es válido pero no se podrán calcular los intervalos de confianza para los valores predichos. Las predicciones de las componentes para el año 2018 se pueden observar en la figura 6.16.

	c1	c2
Independencia	1	1
Media cero	0,0757	0,9765
Normalidad	0,8556	0,3491
Estacionariedad	0,02332218	0,01

Cuadro 6.8: Validación de os modelos ajustados para la primera componente principal dinámica, c1, y para la segunda componente principal dinámica, c2. Se incluyen los p-valores de los contrastes sobre las hipótesis en las que se basa el modelo.

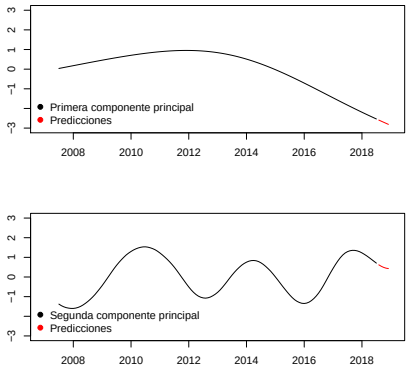


Figura 6.16: Primera componente principal dinámica, en negro, con sus predicciones para el 2018, en rojo, (gráfica superior) y segunda componente principal dinámica, en negro, con sus predicciones para el 2018, en rojo, (gráfica inferior), en Agosto

Capítulo 7

Regresión lineal dinámica

En este capítulo se intentará ajustar un modelo de regresión dinámica para poder predecir los valores futuros del PIB, pero antes de eso, se ajustará un modelo para predecir valores futuros del consumo de hogares gallegos puesto que en la obtención de los datos internos de ABANCA sólo se ha seleccionado el grupo familias. Si el ajuste para la serie consumo es bueno se podrá justificar que nuestras variables explican la evolución del PIB, pues el consumo de los hogares representa, más o menos, un 60 % del valor del PIB. Por lo tanto, la variable respuesta será el PIB o el consumo de hogares gallegos y las variables explicativas serán la primera componente principal dinámica, la segunda componente principal dinámica y el PIB o consumo en hogares gallegos retardados tres instantes de tiempo, respectivamente. Si no se incluyera en el modelo esta última variable, como se observa en la figura 7.1, no se sabría el valor que toma el PIB en alguno de los instantes. Además, dependiendo del instante temporal en el cual nos encontremos vamos a tener más o menos información, por ejemplo, si nos encontramos a principios de Enero del 2018 dispondremos de datos internos hasta Diciembre del 2017 y de valores del PIB y del consumo de hogares gallegos hasta el tercer trimestre del 2017, pues los valores del cuarto trimestre de ese año se publican el 1 de marzo del 2018. Por lo que para predecir los valores del PIB o del consumo para el año 2018, será necesario predecir el valor del PIB en el último trimestre del año 2017. En este caso en el modelo se incluirá el PIB o consumo desde Julio del 2007 a Septiembre de 2017, la primera y segunda componente principal dinámica desde Julio del 2007 a Septiembre de 2017 y el PIB o consumo retardado tres instantes, por lo que esta variable oscilará entre Abril 2007 y Junio 2017. Para la predicción del PIB para el cuarto trimestre del 2017, es decir, para los meses Octubre, Noviembre y Diciembre de dicho año, vamos a disponer de los valores de las componentes desde Octubre a Diciembre del 2017, las cuales son conocidas para nosotros, y del PIB o consumo desde Julio a Septiembre del 2017, los cuales son conocidos. Para la predicción del siguiente trimestre, es decir, de los meses de Enero, Febrero y Marzo del 2018, se necesitarán los valores de las componentes de Enero a Marzo del 2018 y del PIB o consumo retardado tres instantes, es decir, de los valores de los meses de Octubre, Noviembre y Diciembre del año 2017. Estos tres últimos valores ya se han predicho. Para los valores de las componentes principales entre Enero y Marzo del 2018 se necesitará ajustar un modelo para poder obtener proyecciones de estos valores, pues se disponen de datos internos hasta Diciembre del 2017. Para entenderlo mejor se puede observar los cuadros 7.2 y 7.3, en los cuales se describen las variables a utilizar en el modelo dependiendo del instante temporal en el que nos encontremos, y los cuadros 7.4, 7.5 y 7.6, en los cuales se describen los valores que se utilizan para realizar las predicciones. Nótese que la primera y segunda componente principal dinámicas se denotan por C1 y C2 respectivamente, r indica que la variable es conocida en ese instante y p indica que dicha variable es una predicción.

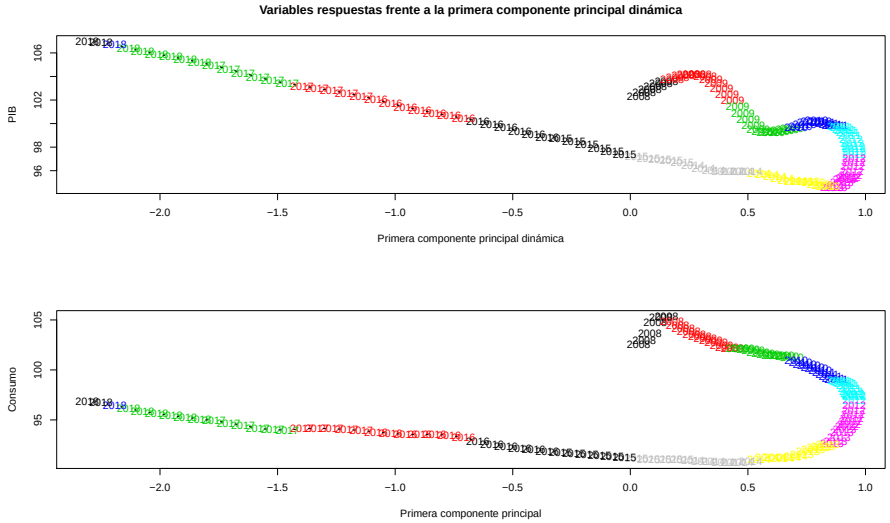


Figura 7.1: Variables respuestas frente a la primera componente principal dinámica, donde cada color representa los distintos años

Supongamos que nos encontramos en Abril del 2018, con lo cual, observando el cuadro 7.2, se podrá realizar una regresión dinámica en la cual la variable respuesta es el PIB desde Julio del 2007 hasta Diciembre del 2017 y las variables explicativas el PIB retardado tres instantes, es decir, desde Abril del 2007 hasta Septiembre del 2017 y la primera y segunda componente principal dinámica desde Julio del 2007 hasta Diciembre del 2017. El cuadro 7.1 muestra que nuestro modelo es válido.

	Regresión dinámica
Independencia	1
Ljung-Box	0,9587
Media cero	0,3176
Normalidad	$2{,}2e-16$
Estacionariedad	0,01

Cuadro 7.1: Validación del modelo de regresión lineal dinámica ajustado. Se incluyen los p-valores de los contrastes sobre las hipótesis en las que se basa el modelo.

Al ser dicho modelo válido podremos obtener predicciones de valores futuros. La figura 7.2 nos muestra los valores observados del PIB en negro, los valores ajustados mediante la regresión dinámica en azul y los valores predichos para el año 2018 en rojo. Como se puede observar en dicha figura las predicciones no muestran pendiente cuando lo lógico, según las opiniones expertas de la entidad, es que mostraran una pendiente positiva.

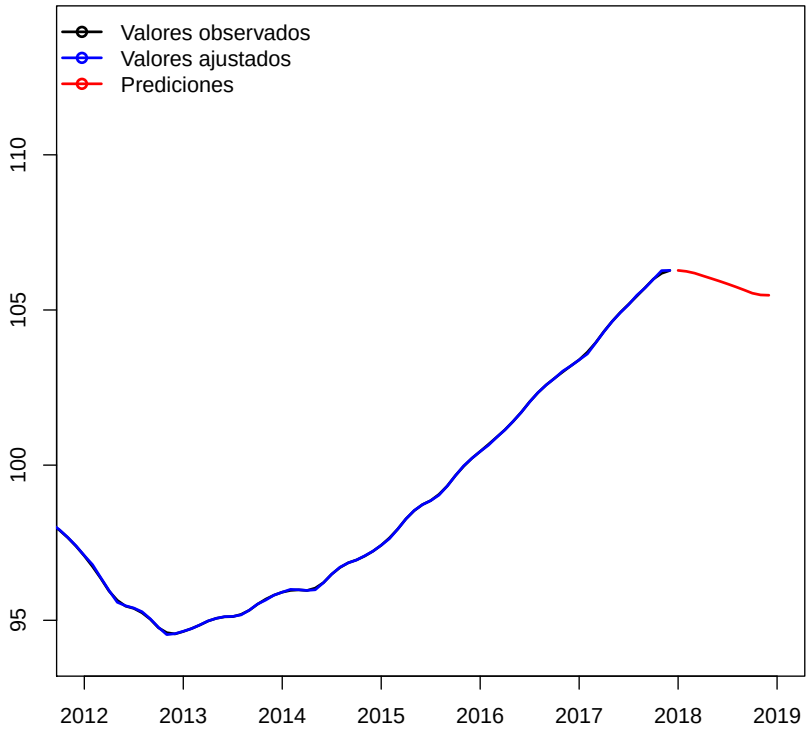


Figura 7.2: Valores observados, valores ajustados y predicciones realizando una regresión dinámica

	Variable respuesta	Variables explicativas
Enero	PIB 7/2007-9/2017	PIB/Consumo retardado tres instantes temporales, 4/2007-6/2017
	o	Primera componente 7/2007-9/2017
	Consumo 7/2007-9/2017	Segunda componente 7/2007-9/2017
Febrero	PIB 7/2007-9/2017	PIB/Consumo retardado tres instantes temporales, 4/2007-6/2017
	o	Primera componente 7/2007-9/2017
	Consumo 7/2007-9/2017	Segunda componente 7/2007-9/2017
Marzo	PIB 7/2007 a 12/2017	PIB/Consumo retardado tres instantes temporales, 4/2007-9/2017
	o	Primera componente de 7/2007-12/2017
	Consumo 7/2007-12/2017	Segunda componente 7/2007-12/2017
Abril	PIB 7/2007-12/2017	PIB/Consumo retardado tres instantes temporales, 4/2007-9/2017
	o	Primera componente 7/2007-12/2017
	Consumo 7/2007-12/2017	Segunda componente 7/2007-12/2017
1-30 de Mayo	PIB 7/2007-12/2017	PIB/Consumo retardado tres instantes temporales, 4/2007-9/2017
	o	Primera componente 7/2007-12/2017
	Consumo desde 7/2007-12/2017	Segunda componente 7/2007-12/2017

Cuadro 7.2: Variables del modelo dependiendo del instante temporal en el cual nos encontremos

	Variable respuesta	Variables explicativas
31 de Mayo	PIB 7/2007-3/2018	PIB/Consumo retardado tres instantes temporales, 4/2007-12/2017
	o	Primera componente del 7/2007-3/2018
	Consumo 7/2007-3/2018	Segunda componente 7/2007-3/2018
Junio	PIB 7/2007-3/2018	PIB/Consumo retardado tres instantes temporales, 4/2007-12/2017
	o	Primera componente 7/2007-3/2018
	Consumo 7/2007-3/2018	Segunda componente 7/2007-3/2018
Julio	PIB 7/2007-3/2018	PIB/Consumo retardado tres instantes temporales, 4/2007-12/2017
	o	Primera componente 7/2007-3/2018
	Consumo 7/2007-3/2018	Segunda componente 7/2007-3/2018
1-29	PIB 7/2007-3/2018	PIB/Consumo retardado tres instantes temporales, 4/2007-12/2017
	o	Primera componente 7/2007-3/2018
Agosto	Consumo 7/2007-3/2018	Segunda componente 7/2007-3/2018

Cuadro 7.3: Variables del modelo dependiendo del instante temporal en el cual nos encontremos

Instante temporal	Q_4^{2017}	Q_1^{2018}	Q_2^{2018}	Q_3^{2018}	Q_4^{2018}
Enero	PIB/Consumo 7/2017-9/2017 (r) C1, C2 11/2017-12/2017 (r)	PIB/Consumo 10/2017-12/2017 (p) C1, C2 1/2018-3/2018 (p)	PIB/Consumo 1/2018-3/2018 (p) C1, C2 4/2018-6/2018 (p)	PIB/Consumo 4/2018-7/2018 (p) C1, C2 7/2018-9/2018 (p)	PIB/Consumo 8/2018-10/2018 (p) C1, C2 10/2018-12/2018 (p)
Febrero	PIB/Consumo 7/2017-9/2017 (r) C1, C2 11/2017-12/2017 (r)	PIB/Consumo 10/2017-12/2017 (p) C1, C2 1/2018-3/2018 (1/2018 r, resto p)	PIB/Consumo 1/2018-3/2018 (p) C1, C2 4/2018-6/2018 (p)	PIB/Consumo 4/2018-7/2018 (p) C1, C2 7/2018-9/2018 (p)	PIB/Consumo 8/2018-10/2018 (p) C1, C2 10/2018-12/2018 (p)
Marzo	Conocido	PIB/Consumo 10/2017-12/2017 (r) C1, C2 1/2018-3/2018 (1/2018-2/2018 r, resto p)	PIB/Consumo 1/2018-3/2018 (p) C1, C2 4/2018-6/2018 (p)	PIB/Consumo 4/2018-7/2018 (p) C1, C2 7/2018-9/2018 (p)	PIB/Consumo 8/2018-10/2018 (p) C1, C2 10/2018-12/2018 (p)
Abril	Conocido	PIB/Consumo 10/2017-12/2017 (r) C1, C2 1/2018-3/2018 (r)	PIB/Consumo 1/2018-3/2018 (p) C1, C2 4/2018-6/2018 (p)	PIB/Consumo 4/2018-7/2018 (p) C1, C2 7/2018-9/2018 (p)	PIB/Consumo 8/2018-10/2018 (p) C1, C2 10/2018-12/2018 (p)

Cuadro 7.4: Valores que tomarían las variables explicativas para obtener una predicción

Instante temporal	Q_4^{2017}	Q_1^{2018}	Q_2^{2018}	Q_3^{2018}	Q_4^{2018}
1-30		PIB/Consumo	PIB/Consumo	PIB/Consumo	PIB/Consumo
	Conocido	10/2017-12/2017 (r)	1/2018-3/2018 (p)	4/2018-7/2018 (p)	8/2018-10/2018 (p)
de Mayo		C1, C2	C1, C2	C1, C2	C1, C2
		1/2018-3/2018 (r)	4/2018-6/2018 (4/2018 r, resto p)	7/2018-9/2018 (p)	10/2018-12/2018 (p)
			PIB/Consumo	PIB/Consumo	PIB/Consumo
	Conocido	Conocido	1/2018-3/2018 (p)	4/2018-7/2018 (p)	8/2018-10/2018 (p)
31 de Mayo			C1, C2	C1, C2	C1, C2
			4/2018-6/2018 (4/2018 r, resto p)	7/2018-9/2018 (p)	10/2018-12/2018 (p)
			PIB/Consumo	PIB/Consumo	PIB/Consumo
	Conocido	Conocido	1/2018-3/2018 (p)	4/2018-7/2018 (p)	8/2018-10/2018 (p)
Junio			C1, C2	C1, C2	C1, C2
			4/2018-6/2018 (4/2018-5/2018 r, resto p)	7/2018-9/2018 (p)	10/2018-12/2018 (p)
			PIB/Consumo	PIB/Consumo	PIB/Consumo
	Conocido	Conocido	1/2018-3/2018 (p)	4/2018-7/2018 (p)	8/2018-10/2018 (p)
Julio			C1, C2	C1, C2	C1, C2
			4/2018-6/2018 (r)	7/2018-9/2018 (p)	10/2018-12/2018 (p)

Cuadro 7.5: Valores que tomarían las variables explicativas para obtener una predicción

Instante temporal	Q_4^{2017}	Q_1^{2018}	Q_2^{2018}	Q_3^{2018}	Q_4^{2018}
1-29 de Agosto	Conocido	Conocido	PIB/Consumo	PIB/Consumo	PIB/Consumo
			1/2018-3/2018 (p)	4/2018-7/2018 (p)	8/2018-10/2018 (p)
			C1, C2	C1, C2	C1, C2
			4/2018-6/2018 (r)	7/2018-9/2018 (7/2018 r, resto p)	10/2018-12/2018 (p)

Cuadro 7.6: Valores que tomarían las variables explicativas para obtener una predicción

Capítulo 8

Modelo GAM

En el capítulo anterior se ha llegado a la conclusión de que la regresión lineal dinámica, en este caso, no es una buena elección para obtener predicciones de valores futuros, aunque el ajuste que realiza dicho modelo es muy bueno. Una causa, puede ser, que los residuos tengan mucho peso en el modelo o que la relación entre las variables explicativas y la variable respuesta no sea lineal, sino que exista una relación más complicada. Por este último motivo, en este capítulo, se propondrá un modelo GAM para obtener las predicciones del consumo de hogares gallegos y del PIB para el año 2018.

En primer lugar se ajustará un modelo GAM para ambos casos y se estudiará si este es válido. Como estamos tratando con series temporales, seguramente no se verifique la hipótesis de independencia. Si esto ocurre, es decir, si el modelo no es válido, se propondrá un modelo ARMA para los residuos y se ajustarán conjuntamente los parámetros de la regresión y del ARMA. Finalmente se chequeará si el modelo es válido, es decir, si pasa los contrastes mencionados en la sección 3.4. Una vez obtenido un modelo válido se obtendrán predicciones de los valores futuros. Al igual que en el capítulo anterior, tendremos que tener en cuenta en qué instante temporal nos encontramos, es decir, se van a tener en cuenta los cuadros 7.2, 7.3, 7.4, 7.5 y 7.6.

Consumo de hogares gallegos

Comenzaremos ajustando un modelo GAM, en el cual, la variable respuesta sea el consumo en hogares gallegos y las variables explicativas sean el consumo en hogares gallegos retardados tres instantes temporales, la primera componente principal dinámica y la segunda componente principal dinámica. A estas dos últimas variables se les aplicarán funciones suavizadoras. Se realizará con detalle uno de los casos. El resto son análogos y se incluirán únicamente los resultados.

Supongamos que nos encontramos a principios de año, es decir, en el mes de Enero. Entonces, siguiendo los cuadros 7.2 y 7.3, la variable respuesta será el consumo desde Julio del 2007 hasta Septiembre del 2017 y las variables explicativas serán el consumo retardado tres instantes temporales, es decir, el consumo desde Abril del 2007 hasta Junio del 2017, y la primera y segunda componente principal desde Julio del 2007 hasta Septiembre del 2017.

El cuadro 8.1 muestra la validación del modelo GAM utilizado. Como se puede observar, si se toma un nivel de significación del 5 %, el modelo no es válido, pues no se verifican los contrastes de independencia y normalidad, ya que los p-valores obtenidos son menores que el nivel de significación considerado. Además, al observar la fas y fap de dicho modelo, figura 8.1, se llega a la conclusión de que los residuos no son ruido blanco, con lo cual se propondrá un modelo $ARMA$ para los residuos y se ajustarán conjuntamente los parámetros de la regresión y del $ARMA$ de los residuos de dicho modelo. En este caso, llegaría con ajustar un $AR(3)$.

	Modelo GAM
Test de Ljung-Box	$2{,}2e-16$
Media cero	0,9478
Normalidad	$2{,}2e-16$
Estacionariedad	0,01

Cuadro 8.1: Validación del modelo GAM. Se incluyen los p-valores de los contrastes sobre las hipótesis en las que se basa el modelo.

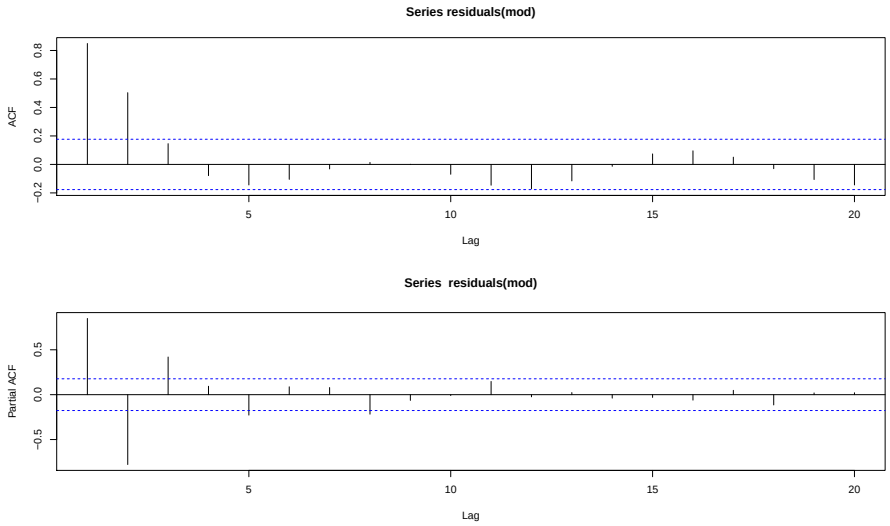


Figura 8.1: fas y fap de los residuos del modelo GAM

Una vez ajustados conjuntamente los parámetros de la regresión y del modelo $ARMA$ propuesto para los residuos, se realiza un chequeo del modelo. Esta validación se muestra en el cuadro 8.2.

En dicho cuadro se puede observar que, con un nivel de significación del 5 %, el modelo es válido, pues los p-valores de los contrastes de independencia y media cero son mayores que dicho nivel de significación, por lo que no hay evidencias para rechazar la hipótesis nula, y el p-valor del contraste de estacionariedad es menor que el nivel de significación, por lo que se rechazará la hipótesis nula de estacionariedad. Sin embargo, el p-valor del contraste de normalidad es menor que el nivel de significación considerado. Esto último no invalida el modelo, lo que indica es que no se pueden obtener los intervalos de confianza para las predicciones.

Por otro lado, observando las fas y fap de los residuos del modelo, figura 8.2, se podría afirmar que los residuos de dicho modelo no son ruido blanco, de modo que el modelo no es válido. Sin embargo, en este caso, no es así, pues al desagregar la serie del consumo, es decir, al pasar de una frecuencia trimestral a una frecuencia mensual se está añadiendo ruido, el cual, no se puede explicar. Para comprobar que esta última afirmación es cierta se realizará una pequeña simulación (véase la simulación 8.0.1 más abajo).

	Regresión dinámica
Test de Ljung-Box	0,1807
Media cero	0,6295
Normalidad	$2{,}2e-16$
Estacionariedad	0,01

Cuadro 8.2: Validación del modelo GAM de los residuos del modelo que ajusta conjuntamente los parámetros de la regresión y del modelo $ARMA$ propuesto para los residuos. Se incluyen los p-valores de los contrastes sobre las hipótesis en las que se basa el modelo.

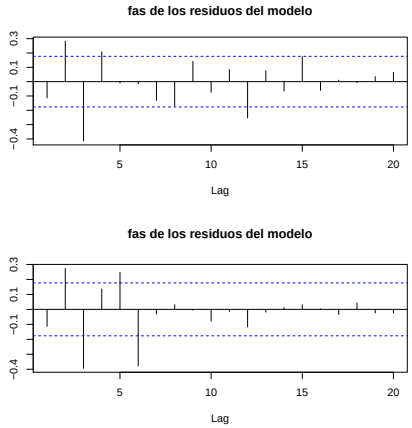
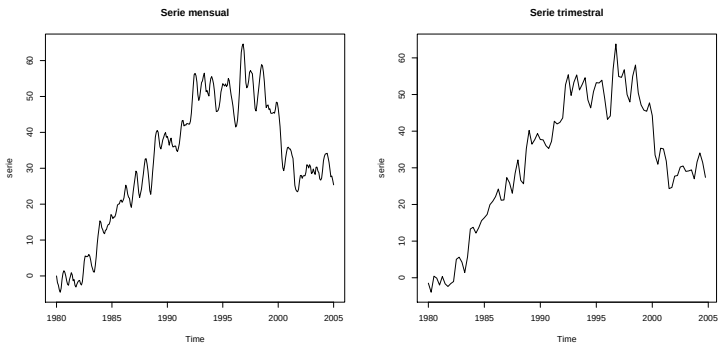


Figura 8.2: fas y fap de los residuos del modelo que ajusta conjuntamente los parámetros de la regresión y del modelo $ARMA$ propuesto para los residuos

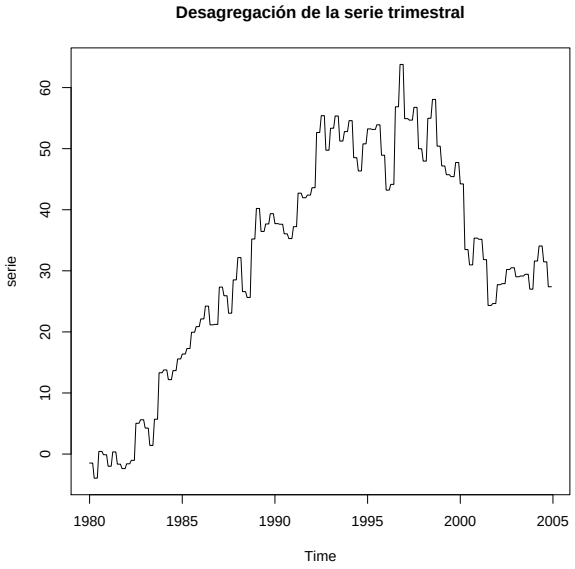
Simulación 8.0.1. *En esta simulación se intentará probar que en el proceso de desagregación de una serie, en el cual se pasa de una frecuencia trimestral a una mensual, se está añadiendo ruido, el cual, no se va a poder explicar. Para ello se simulará una serie con frecuencia mensual, sabiendo, que sigue un modelo* $ARIMA(4,1,0)\times(0,0,0)$ *y se trimestralizará. La serie con frecuencia mensual se muestra en la figura 8.3a y su trimestralización se muestra en la figura 8.3b.*



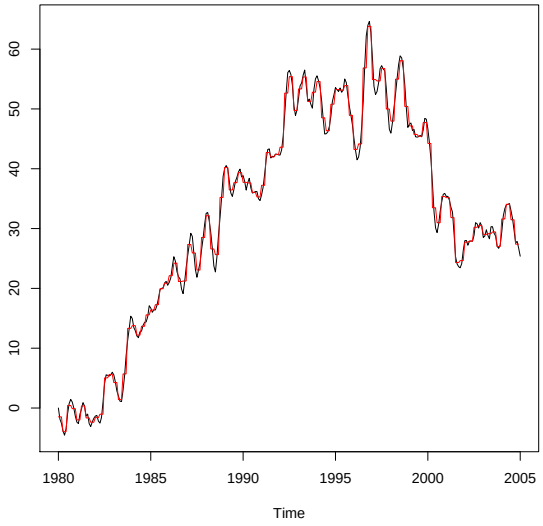
(a) Serie mensual simulada (b) Serie trimestralizada

Figura 8.3: Series simuladas

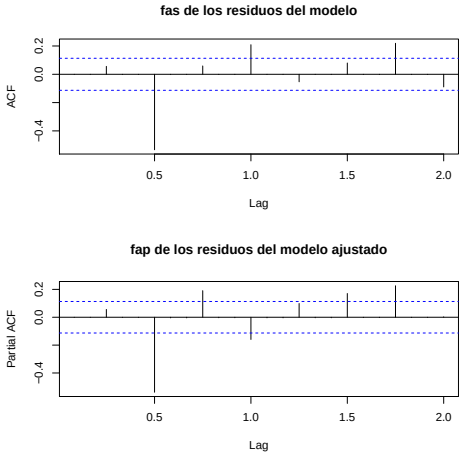
Una vez obtenida la trimestralización de la serie se desagregará, es decir, se volverá a la serie con frecuencia mensual, la cual se debería poder ajustar con un modelo $ARIMA(4,1,0) \times (0,0,0)$*, pues deberíamos haber obtenido la serie original, es decir, deberíamos haber obtenido la serie generada mediante el proceso* $ARIMA(4,1,0) \times (0,0,0)$*. La figura 8.4a muestra la desagregación de la serie trimestral y la figura 8.4c muestra la fas y fap de los residuos del modelo ajustado. Como se puede observar en esta última, los residuos de dicho modelo no son ruido blanco. Por lo que se puede concluir que en el proceso de desagregación se está añadiendo "algo" que no se puede explicar. Sin embargo, observando el cuadro 8.3, se puede asegurar que el modelo pasa los contrastes de validación.*



(a) Desagregación de la serie trimestral



(b) Serie mensual original en negro vs serie desagregada en rojo



(c) fas y fap de los residuos del modelo

Figura 8.4: Desagregación de la serie trimestral, fas y fap de los residuos del modelo ajustado

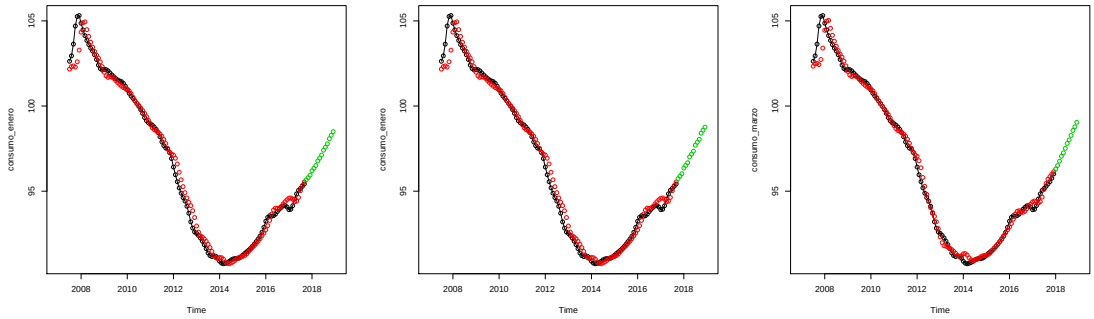
	Modelo ajustado
Test de Ljung-Box	0,9806
Media cero	0,5198
Normalidad	$2{,}2e-16$
Estacionariedad	0,01

Cuadro 8.3: Validación del modelo. Se incluyen los p-valores de los contrastes sobre las hipótesis en las que se basa el modelo.

Una vez validado el modelo seleccionado se podrá obtener predicciones de valores futuros. Para ello se seguirá lo explicado en los cuadros 7.4, 7.5 y 7.6. Estas predicciones se muestran en los cuadros 8.4, 8.5 y 8.6, donde se ha denotado por:

- Q_i^t al consumo de hogares gallegos en el trimestre i del año t, es decir, Q_4^{2017} denota el valor del consumo de hogares gallegos en el cuarto trimestre del año 2017.
- t/i la variación del consumo de hogares gallegos en el trimestre i del año t, es decir, 2017/4 es la variación del consumo de hogares gallegos en el cuarto trimestre del año 2017.
- t la variación interanunal del año t del consumo de hogares gallegos, es decir, 2017 denotará la variación interanual de dicho año.

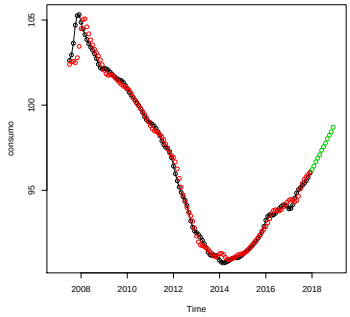
Los cuadros 8.4, 8.5 y 8.6 también muestran las distintas predicciones para el consumo de hogares gallegos en el resto de instantes temporales en los que uno se puede encontrar. Además, en la figura 8.5 se pueden observar los valores conocidos del consumo en hogares gallegos en negro, los valores ajustados en rojo y las predicciones en verde.



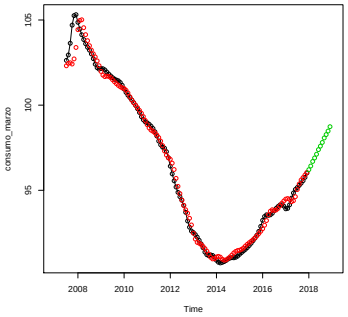
(a) Enero

(b) Febrero

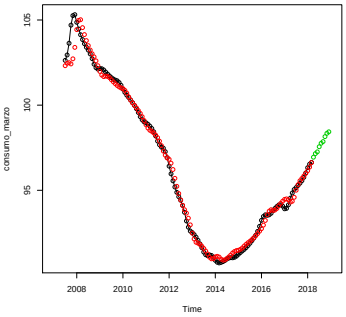
(c) Marzo



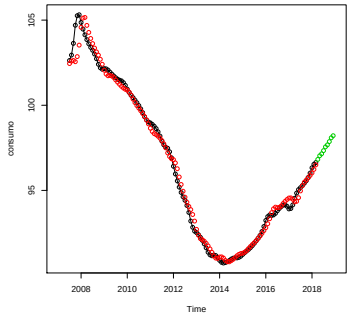
(d) Abril



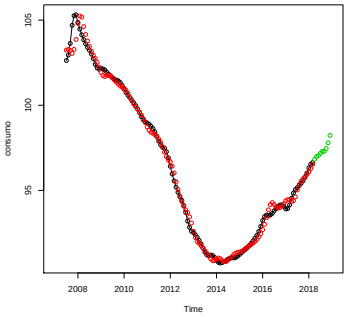
(e) Del 1 al 30 de Mayo



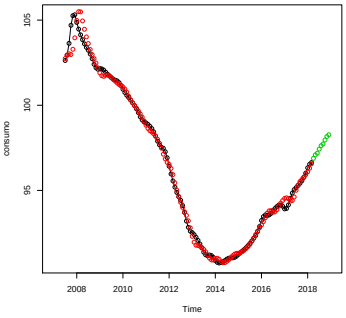
(f) 31 de Mayo



(g) Junio



(h) Julio



(i) Agosto

Figura 8.5: Valores observados del consumo de hogares gallegos en negro, valores ajustados en rojo y predicciones en verde

		Q_4^{2017}	2017/4	2017	Q_1^{2018}	2018/1	Q_2^{2018}	2018/2	Q_3^{2018}	2018/3	Q_4^{2018}	2018/4	2018
Enero	Predicción	95,79862	1,783488	1,295872	96,34524	2,494937	96,94174	2,248430	97,58725	2,410795	98,15524	2,459978	2,403535
	Valor real	95,78	1,8	1,3	96,50	2,7	96,81	2,1	-	-	-	-	-
Febrero	Predicción	95,88071	1,870705	1,317676	96,51377	2,674222	97,18122	2,501018	97,87393	2,711645	98,44981	2,679478	2,641591
	Valor real	95,78	1,5	1,8	96,50	2,7	96,81	2,1	-	-	-	-	-
Marzo	Predicción	conocido	conocido	conocido	96,52211	2,683100	97,27132	2,596052	98,02919	2,874587	98,79769	3,150646	2,826096
	Valor real				96,50	2,7	96,81	2,1	-	-	-	-	-
Abril	Predicción	conocido	conocido	conocido	96,44055	2,596334	97,10883	2,424675	97,78272	2,615935	98,45742	2,795383	2,608082
	Valor real				96,50	2,7	96,81	2,1	-	-	-	-	-

Cuadro 8.4: Predicción del consumo en hogares gallegos

		Q_4^{2017}	2017/4	2017	Q_1^{2018}	2018/1	Q_2^{2018}	2018/2	Q_3^{2018}	2018/3	Q_4^{2018}	2018/4	2018
1-30	Predicción				96,44373	2,599717	97,13501	2,452289	97,8253	2,660614	98,49549	2,835136	2,636939
		conocido	conocido	conocido									
de Mayo	Valor real				96,50	2,7	96,81	2,1	-	-	-	-	-
	Predicción						97,1174	2,433711	97,72706	2,557515	98,31156	2,643096	2,573474
31 de mayo		conocido	conocido	conocido	conocido	conocido							
	Valor real						96,81	2,1	-	-	-	-	-
	Predicción						97,00144	2,311397	97,52287	2,343236	98,05727	2,343236	2,422954
Junio		conocido	conocido	conocido	conocido	conocido							
	Valor real						96,81	2,1	-	-	-	-	-
	Predicción						96,94968	2,256804	97,25441	2,061506	97,82725	2,137450	2,278834
Julio		conocido	conocido	conocido	conocido	conocido							
	Valor real						96,81	2,1	-	-	-	-	-

Cuadro 8.5: Predicción del consumo en hogares gallegos

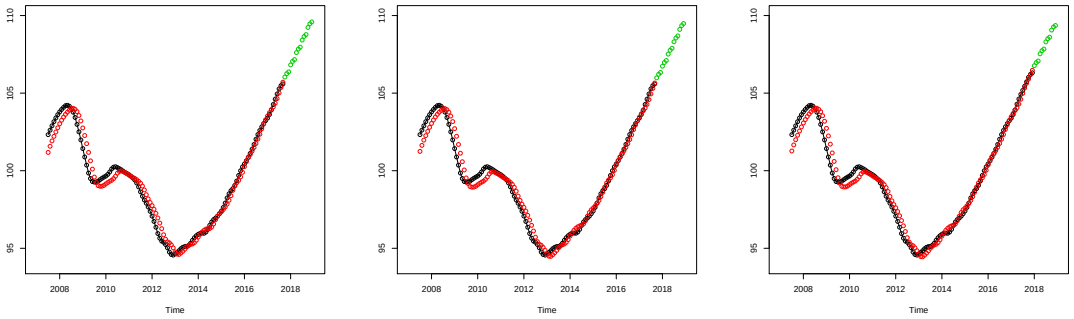
		Q_4^{2017}	2017/4	2017	Q_1^{2018}	2018/1	Q_2^{2018}	2018/2	Q_3^{2018}	2018/3	Q_4^{2018}	2018/4	2018
1-29	Predicción						97,05311	2,365905	97,59303	2,416860	98,1322	2,455833	2,474543
		conocido	conocido	conocido	conocido	conocido							
Agosto	Valor real						96,81	2,1	-	-	-	-	-

Cuadro 8.6: Predicción del consumo en hogares gallegos

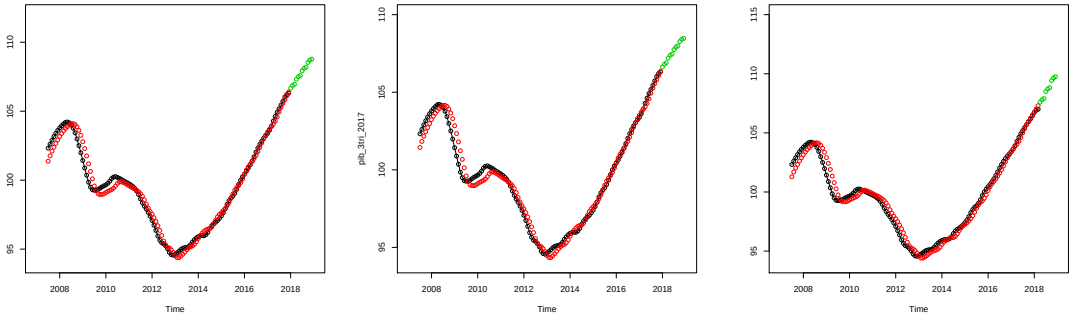
PIB gallego

El modelo obtenido anteriormente proporciona unas predicciones aceptables sobre el consumo de hogares gallegos. Por lo tanto, se puede justificar que las variables explicativas pueden explicar el comportamiento del PIB, puesto que el consumo de hogares gallegos representa, más o menos, un 60 % del valor del PIB.

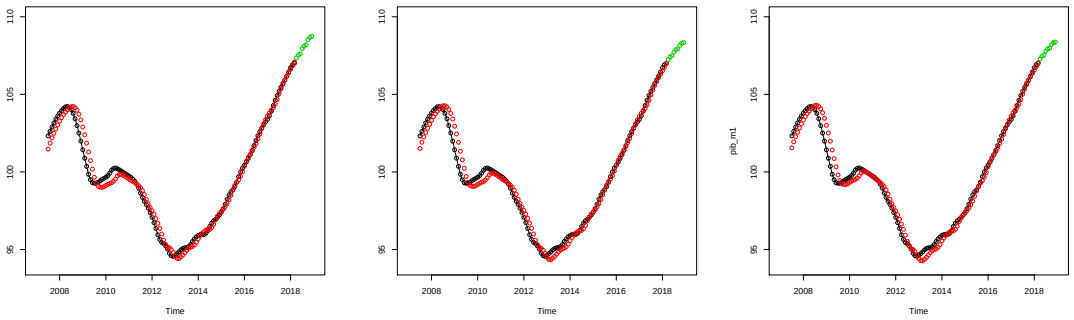
Una vez obtenido un modelo válido, a través del proceso descrito al principio de este capítulo, podremos obtener predicciones del PIB. Dichas predicciones se muestran en los cuadros 8.7 y 8.8. Además, en la figura 8.5 se puede observar los valores observados del consumo en hogares gallegos en negro, los valores ajustados en rojo y las predicciones en verde.



(a) Enero (b) Febrero (c) Marzo



(d) Abril (e) Del 1 al 30 de Mayo (f) 31 de Mayo



(g) Junio (h) Julio (i) Agosto

Figura 8.6: Valores observados del consumo de hogares gallegos en negro, valores ajustados en rojo y predicciones en verde

		Q_4^{2017}	2017/4	2017	Q_1^{2018}	2018/1	Q_2^{2018}	2018/2	Q_3^{2018}	2018/3	Q_4^{2018}	2018/4	2018
Enero	Predicción	106,2205	3,096651	3,099163	107,0068	3,248565	107,8058	3,054947	108,6146	3,001061	109,4305	3,022082	3,081664
	Valor real	106,18	3,1	3,1	106,88	3,1	107,50	2,8	-	-	-	-	-
Febrero	Predicción	106,1777	3,055175	3,088794	106,9387	3,182836	107,723	2,975851	108,5186	2,909965	109,3147	2,954451	3,005776
	Valor real	106,18	3,1	3,1	106,88	3,1	107,50	2,8	-	-	-	-	-
Marzo	Predicción	Conocido	Conocido	Conocido	106,936	3,180268	107,7035	2,957209	108,4711	2,864937	109,2299	2,872402	2,968704
	Valor real				106,88	3,1	107,50	2,8	-	-	-	-	-
Abril	Predicción	Conocido	Conocido	Conocido	106,8306	3,078520	107,4655	2,729689	108,0792	2,493292	108,6701	2,345157	2,661665
	Valor real				106,88	3,1	107,50	2,8	-	-	-	-	-
1-30 de Mayo	Predicción	Conocido	Conocido	Conocido	106,7791	3,028805	107,3477	2,617019	107,8826	2,306855	108,3839	2,075648	2,502282
	Valor real				106,88	3,1	107,50	2,8	-	-	-	-	

Cuadro 8.7: Predicción del PIB gallego

		Q_4^{2017}	2017/4	2017	Q_1^{2018}	2018/1	Q_2^{2018}	2018/2	Q_3^{2018}	2018/3	Q_4^{2018}	2018/4	2018
	Predicción						107,7664	3,017347	108,6784	3,061554	109,6058	3,22638	3,107872
31 de Mayo		Conocido	Conocido	Conocido	Conocido	Conocido							
	Valor real						107,50	2,8	-	-	-	-	-
	Predicción						107,4979	2,760633	108,0818	2,495809	108,6429	2,319566	2,675554
Junio		Conocido	Conocido	Conocido	Conocido	Conocido							
	Valor real						107,50	2,8	-	-	-	-	-
	Predicción						107,3882	2,655797	107,8477	2,273767	108,2627	1,961437	2,504302
Julio		Conocido	Conocido	Conocido	Conocido	Conocido							
	Valor real						107,50	2,8	-	-	-	-	-
1-29	Predicción						107,4221	2,688134	107,8965	2,320034	108,3096	2,005614	2,534997
		Conocido	Conocido	Conocido	Conocido	Conocido							
de Agosto	Valor real						107,50	2,8	-	-	-	-	-

Cuadro 8.8: Predicción del PIB gallego

Capítulo 9

Obtención de un semáforo que refleje el comportamiento del consumo de hogares gallegos y del PIB gallego en cada una de las provincias

En el capítulo anterior se ha obtenido un modelo para predecir el consumo de hogares gallego y otro para predecir el PIB gallego. Recordemos que en ambos, aparte de las componentes principales dinámicas, se utilizaban tanto el consumo de hogares gallegos como el PIB retardado tres instantes como variables explicativas, y como variable respuesta el consumo de hogares gallegos o el PIB gallego. En este capítulo, se intentará construir un semáforo que nos indique cómo evolucionan las dos variables de interés en cada una de las provincias gallegas, extrapolando, para ello, los modelos obtenidos en el capítulo anterior. Sin embargo, en este caso, el modelo GAM no tiene porqué ser válido, pues lo que nos interesa es que explique el comportamiento de las variables a estudiar. Por ese motivo vamos a considerar los modelos anteriores pero, en los cuales, se suprimirá como variables explicativa tanto el consumo en hogares gallegos como el PIB retardado tres instantes. Para extrapolar estos modelos al nivel provincial, tendremos que hacer lo siguiente:

1 "Reconstruir" las componentes dinámicas para cada provincia gallega. En el modelo se incluían variables que se conocen a nivel de Galicia pero no a nivel provincial. Por ese motivo es necesario relacionar las componentes principales dinámicas gallegas con las variables internas referida únicamente a la provincia de interés. En este caso vamos a suponer que sólo conocemos los valores de las variables hasta Marzo del 2018. Se ajustará un modelo de regresión lineal simple, donde la variable respuesta será la componente principal dinámica gallega y las variables explicativas las series internas de la entidad referidas a la comunidad autónoma. De esta regresión se obtendrán unos coeficientes asociados a cada variable interna. Por lo que, al multiplicar cada coeficiente por la variable interna de la provincia, las cuales has sido corregidas previamente, se obtendrá una "reconstrucción" de la componente principal dinámica. Es decir, si los coeficientes obtenidos en la regresión lineal son: $c_1, c_2, \cdots, c_{13}$, la "reconstrucción" de la componente principal dinámica será: $c_1 \times v_1 + c_2 \times v_2 + \ldots + c_{13} \times v_{13}$, donde $v_1, \ldots, v_{13}$ representa cada una de las variables referidas a la provincia.

2 Obtener las funciones suavizadoras que se han utilizado en el modelo a nivel Galicia y aplicárselas a la "reconstrucción" de las componentes principales dinámicas, a las que llamaremos B.

3 Obtener los coeficientes del modelo GAM a nivel Galicia, c.

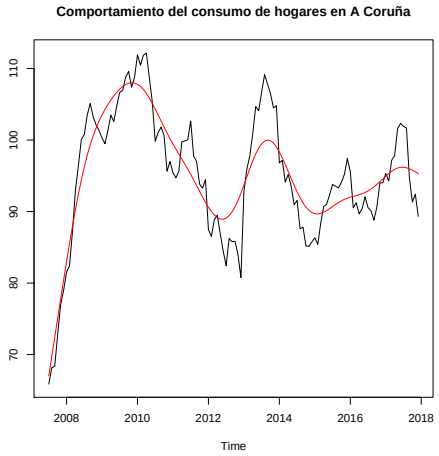
4 El ajuste para cada provincia será: Bc, siendo B la matriz de suavización y c el vector de coeficientes.

Esta será una primera estimación del comportamiento de ambas variables a nivel provincial. La figura

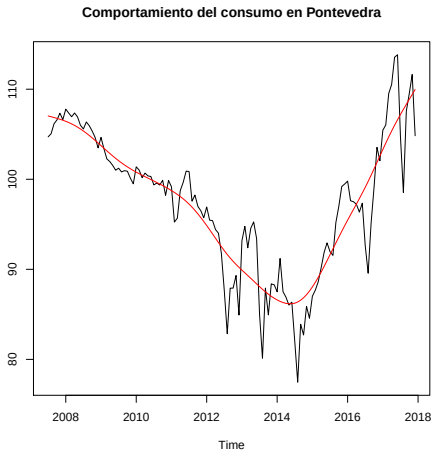
9.1 muestra el comportamiento del consumo de hogares gallegos para cada provincia y la figura 9.2 nos muestra el comportamiento del PIB a nivel provincial.

La figura 9.1 muestra el comportamiento del consumo de hogares gallegos para cada provincia. Dicha figura se puede interpretar de la siguiente manera: el consumo en las provincias de La Coruña, Lugo y Ourense disminuyó en el año 2017, sin embargo, en la provincia de Pontevedra el consumo aumentó en dicho año. La figura 9.2 muestra el comportamiento del PIB a nivel provincial y tiene una interpretación análoga a la anterior.

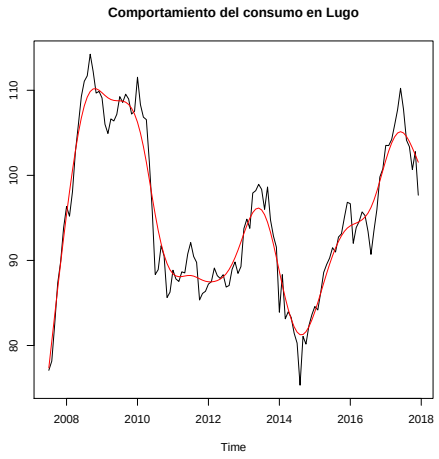
Está en desarrollo la búsqueda de algún elemento de validación para las conclusiones obtenidas con esta desagregación territorial del modelo.



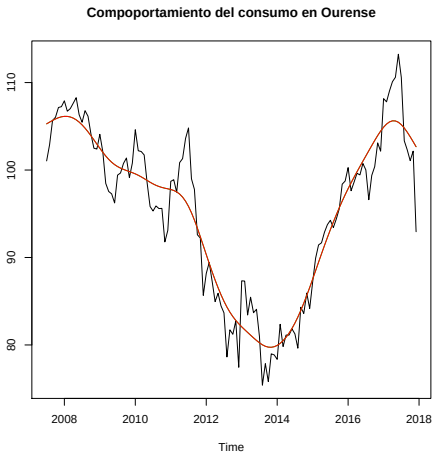
(a) Comportamiento del consumo en A Coruña



(b) Comportamiento del consumo en Pontevedra

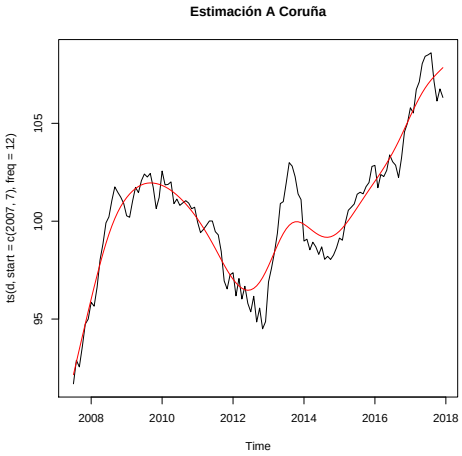


(c) Comportamiento del consumo en Lugo

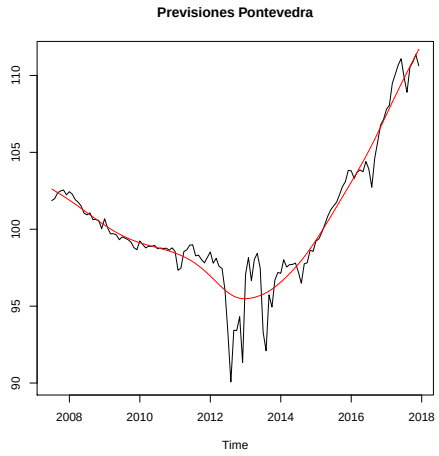


(d) Comportamiento del consumo en Ourense

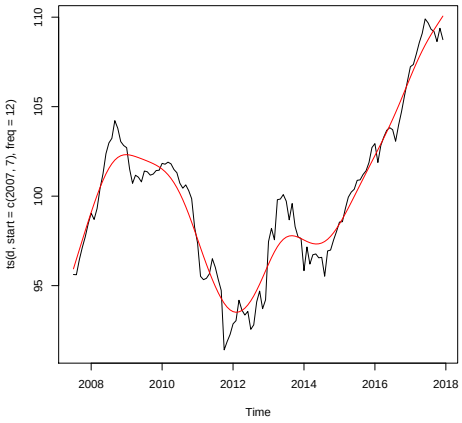
Figura 9.1: Primera estimación del comportamiento del consumo en cada una de las provincias en negro y en rojo una suavización spline



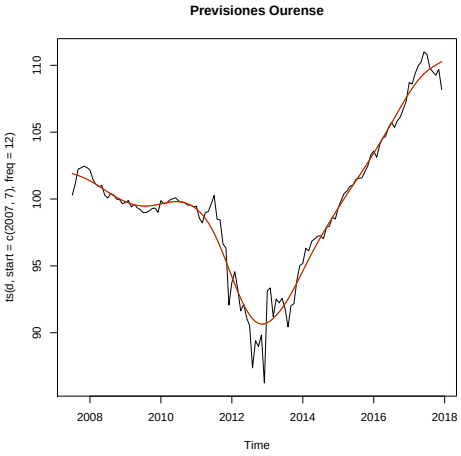
(a) Comportamiento de PIB en A Coruña



(b) Comportamiento de PIB en Pontevedra



(c) Comportamiento de PIB en Lugo



(d) Comportamiento de PIB en Ourense

Figura 9.2: Primera estimación del comportamiento del PIB en cada una de las provincias en negro y en rojo una suavización spline

Capítulo 10

Conclusiones

La finalidad de este trabajo era construir un modelo estadístico que permita anticipar los valores del PIB a partir de información interna de ABANCA. Para ello, se ha manejado un conjunto de 13 variables con información de la entidad financiera. Cada una de las 13 variables analizadas es, en sí misma, una serie temporal, por lo que era de gran interés reducir la dimensión. Para ello, como se ha visto, se utilizaba el análisis de componentes principales dinámicas. Además, todas ellas hacen referencia al segmento familias, por lo que se ha desarrollado un modelo que permite pronosticar la componente de consumo de los hogares del PIB. Una vez desarrollado este modelo, y dado que la componente de consumo de los hogares ha supuesto, históricamente, el 60 % de la variación del PIB, se ha analizado también un modelo que permite pronosticar el PIB gallego a partir de la información interna de la entidad.

Para obtener las proyecciones tanto del PIB como del consumo de hogares gallegos se han ajustado distintos modelos. Como primera opción se tomó la regresión lineal dinámica. Tal y como se ha visto en ese capítulo, aunque el ajuste del modelo parecía bueno, las previsiones dadas por el mismo no parecen encajar con la dinámica actual de la economía gallega. Como segunda opción, se prueba un ajuste no paramétrico, un modelo GAM, que mantiene buenos resultados en el ajuste y un pronóstico más acorde a las opiniones de los expertos. Por último, se intenta extrapolar el modelo desarrollado para Galicia, a cada una de las provincias. Esta aproximación, aunque aún en desarrollo, resulta de gran interés, ya que abre la posibilidad de obtener una medida de la evolución de la economía gallega desagregada por provincias y construida a partir de la información de la entidad. Hecho, este último, que un gran dinamismo al análisis dada la disponibilidad de los datos.

Bibliografía

[1] Arnau Gras J (2001), *Diseños de series temporales: técnicas de análisis*. Edicións Universitat de Barcelona

[2] Box GEP, and Cox DR (1964), *An analysis of transformations*. Journal od the Royal Statistical Society. Serie B, Vol.26, No. 2, 211-252.

[3] Brillinger DR (2001) *Time series: Data Analysis and theory*. SIAM, San Francisco.

[4] CRAN documentation on the R package 'gdpc', https://cran.r-project.org/web/packages/gdpc/gdpc.pdf

[5] CRAN documentation on the R package 'mgcv', https://cran.r-project.org/web/packages/mgcv/mgcv.pdf

[6] CRAN documentation on the R package 'tempdisagg', https://cran.r-project.org/web/packages/tempdisagg/tempdisagg.pdf

[7] Ku WF, Storer RH, Georgakis C (1995), *Disturbance detection and isolation by dynamic principal component analysis*, Chemometrics and intelligent laboratory systems, Vol. 30, 179-196.

[8] Monchón F, Beker VA (2008), *Economía: Principios y aplicaciones*. MC Graw Hill, México, D. F.

[9] Müller, M. (2004), Generalized Linear Models. In: J. Gentle, W. Härdle, Y. Mori (eds): Handbook of Computational Statistics (Volume I). *Concepts and Fundamentals*, Springer-Verlag, Heidelberg, 2004.

[10] Peña D (2002), *Análisis de datos multivariantes*. S.A. MCGRAW-HILL.

[11] Peña D (2010), *Análisis de series temporales*. Alianza Editorial, Madrid.

[12] Peña D, and Yohai VJ (2016), *Generalized Dynamic Principal Components*. Journal of the American Statistical Association, 111, 1121-1131.

[13] Shumway RH, and Stoffer DS (2006), *Time Series Analysis and Its Applications With R Examples*. Springer.

[14] Wood SN, (2006), *Generalized Additive Models: an introduction with R*. CRC Press Taylor & Francis Group.