Trabajo Fin de Máster

Modelización de la actividad económica española en tiempo real Un enfoque de Nowcasting mediante Modelos de Factores Dinámicos

Yago Núñez García

Máster en Técnicas Estadísticas

Curso 2025-2026

Propuesta de Trabajo Fin de Máster

Título en galego: Modelización da actividade económica española en tempo real: un enfoque de Nowcasting mediante Modelos de Factores Dinámicos

Título en español: Modelización de la actividad económica española en tiempo real: Un enfoque de Nowcasting mediante Modelos de Factores Dinámicos

English title: Modeling Spanish Economic Activity in Real Time: A Nowcasting Approach Using Dynamic Factor Models

Modalidad: Modalidad B

Autor: Yago Núñez García, Universidad de La Coruña

Directores: Rubén Fernández Casal, Universidad de La Coruña; Germán Aneiros Pérez, Universidad de La Coruña

Tutores: Teresa Veiga Rodríguez, ABANCA; Ana Blanco Bugueiro, ABANCA; Sergio Díaz Canosa, ABANCA

Breve resumen del trabajo:

En el área de Planificación Estratégica y PMO de ABANCA se lleva a cabo el análisis y seguimiento del entorno macroeconómico de forma periódica. El principal agregado macroeconómico de la economía española, el Producto Interior Bruto, sufre un desfase en la publicación del dato oficial de 30 días desde la finalización del trimestre de referencia. Ante esta problemática, y con el fin de obtener una estimación temprana del pulso de la economía española, se plantea utilizar la metodología del Modelo Factorial Dinámico expresado en el Espacio de Estados, con el fin de explotar el elevado número de series de tiempo macroeconómicas de frecuencia mensual que afectan al cálculo del PIB. Se pretende obtener un factor común que capture el comovimiento de las series del modelo y así obtener una estimación del pulso de la economía española semanas antes de la publicación del dato oficial en el Instituto Nacional de Estadística, que proporcionará información valiosa para la toma de decisiones estratégicas en ABANCA.

Don Rubén Fernández Casal, Codirector de la Universidad de La Coruña, don Germán Aneiros Pérez, Codirector de la Universidad de La Coruña, doña Teresa Veiga Rodríguez, Tutora de ABANCA, doña Ana Blanco Bugueiro, Tutora de ABANCA y don Sergio Díaz Canosa, Tutor de ABANCA, informan que el Trabajo Fin de Máster titulado

Modelización de la actividad económica española en tiempo real: Un enfoque de Nowcasting mediante Modelos de Factores Dinámicos

fue realizado bajo su dirección por don Yago Núñez García 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 15 de Julio de 2026

El director:

Firmado digitalmente por FERNANDEZ CASAL RUBEN - 33317724Q
Fecha: 2026.07.17 22:28:43 +02'00'

Don Rubén Fernández Casal

El director:

ANEIROS PEREZ GERMAN - 76409696Q

Firmado digitalmente por ANEIROS PEREZ GERMAN - 76409696Q
Fecha: 2026.07.17 22:35:41 +02'00'

Don Germán Aneiros Pérez

La tutora:

Doña Teresa Veiga Rodríguez

La tutora:

Doña Ana Blanco Bugueiro

El tutor:

Don Sergio Díaz Canosa

El autor:

NUÑEZ GARCIA YAGO - 58004904P

Firmado digitalmente por NUÑEZ GARCIA YAGO - 58004904P
Fecha: 2026.07.17 22:41:50 +02'00'

Don Yago Núñez García

Declaración responsable.

Para dar cumplimiento a la Ley 3/2022, de 24 de febrero, de convivencia universitaria, referente al plagio en el Trabajo Fin de Máster (Artículo 11, Disposición 2978 del BOE núm. 48 de 2022), **el autor declara** que el Trabajo Fin de Máster presentado es un documento original en el que se han tenido en cuenta las siguientes consideraciones relativas al uso de material de apoyo desarrollado por otros/as autores/as:

- Todas las fuentes usadas para la elaboración de este trabajo han sido citadas convenientemente (libros, artículos, apuntes de profesorado, páginas web, programas, etc.).
- Cualquier contenido copiado o traducido textualmente se ha puesto entre comillas, citando su procedencia.
- Se ha hecho constar explícitamente cuando un capítulo, sección, demostración, etc., sea una adaptación casi literal de alguna fuente existente.

Y acepta que, si se demostrara lo contrario, se le apliquen las medidas disciplinarias que correspondan.

"All models are wrong, but some are useful."
— Box (1976)

Agradecimientos

En primer lugar, quería agradecer a la entidad ABANCA por darme esta oportunidad dentro del Departamento de Planificación Estratégica y PMO para desarrollar el Trabajo de Fin de Máster en la empresa y poder aplicar los conocimientos que he ido adquiriendo a lo largo de este año y medio en el Máster de Técnicas Estadísticas e Investigación Operativa, pudiendo observar de primera mano las dinámicas del mundo laboral y las problemáticas a las que uno se encuentra en la aplicación empírica de los modelos estadísticos a datos reales. En particular, quería agradecer a todos los compañeros con los que he tenido el placer de relacionarme durante mi estancia en la entidad, con especial mención a Teresa Veiga Rodríguez, Ana Blanco Bugueiro y Sergio Díaz Canosa, por su inestimable apoyo y conocimientos aportados, además de ser unas magníficas personas que me han acogido con los brazos abiertos dentro de su equipo.

Por otro lado, agradecer al profesorado del Máster en Técnicas Estadísticas e Investigación Operativa por los conocimientos aportados y, en especial, a mis tutores académicos, Germán Aneiros Pérez y Rubén Fernández Casal.

Índice general

Índice de figuras

Índice de tablas

Resumen

Resumen en español

Anticipar los puntos de giro del ciclo económico es una tarea prioritaria para los equipos de análisis y planificación en el sector bancario. Sin embargo, la medición oficial de la actividad agregada a través del Producto Interior Bruto (PIB) adolece de un notable desfase temporal en su publicación. Esta falta de sincronía con la toma de decisiones en tiempo real motiva el desarrollo de este trabajo, cuyo fin es diseñar un indicador adelantado de alta frecuencia enfocado en el nowcasting de la economía española.

La herramienta propuesta se fundamenta en un Modelo Factorial Dinámico en el espacio de estados, una metodología que reduce la alta dimensionalidad de un gran panel de indicadores mensuales y trimestrales al extraer un único factor común inobservable. Previo a la estimación conjunta mediante el algoritmo de Esperanza-Maximización y el Filtro de Kalman, las series brutas se someten a un tratamiento de desestacionalización, linearización y estabilización de la varianza. Además, se lleva a cabo una criba de variables en búsqueda del panel óptimo sumado a distintos tratamientos del periodo de alta volatilidad generado por la COVID-19.

La validación extramuestral en pseudo-tiempo real demuestra la consistencia del indicador frente a la inercia histórica y ante las previsiones de organizaciones con modelos de referencia, ofreciendo además una alta interpretabilidad económica mediante el gráfico de impacto marginal de cada dato nuevo, junto a esto se realiza un análisis de sensibilidad para estresar el modelo ante situaciones poco probables. Finalmente, el modelo se integra en una plataforma web interactiva construida en R Shiny. Este cuadro de mando, potenciado con interfaces reactivas y un asistente de lenguaje natural en entorno local, democratiza la explotación del indicador dentro de la organización al permitir su uso directo en informes estratégicos sin requerir conocimientos técnicos de programación.

Resumo en galego

Anticipar os puntos de xiro do ciclo económico é unha tarefa prioritaria para os equipos de análise e planificación no sector bancario. Porén, a medición oficial da actividade agregada a través do Produto Interior Bruto (PIB) padece un notable desfasamento temporal na súa publicación. Esta falta de sincronía coa toma de decisións en tempo real motiva o desenvolvemento deste traballo, co fin de deseñar un indicador adiantado de alta frecuencia enfocado no nowcasting da economía española.

A ferramenta proposta fundaméntase nun Modelo Factorial Dinámico no espazo de estados, unha metodoloxía que reduce a alta dimensionalidade dun gran panel de indicadores mensuais e trimestrais ao extraer un único factor común inobservable. Previo á estimación conxunta mediante o algoritmo de Esperanza-Maximización e o Filtro de Kalman, as series brutas sométense a un tratamento de desestacionalización, linearización e estabilización da varianza. Ademais, lévase a cabo un proceso de selección de variables para identificar o panel óptimo, xunto con diversos tratamentos para abordar o período de alta volatilidade provocado pola COVID-19.

A validación en novas mostras en tempo real simulado demostra a consistencia do indicador respecto á inercia histórica e ás previsións de organizacións que empregan modelos de referencia, ao tempo que ofrece unha elevada interpretabilidade económica mediante un gráfico de impacto marxinal para cada

novo punto de datos; ademais, realízase unha análise de sensibilidade para someter o modelo a probas de estrés fronte a escenarios de baixa probabilidade. Finalmente, el modelo intégrase nunha plataforma web interactiva construída en R Shiny. Este cadro de mando, potenciado con interfaces reactivas e un asistente de linguaxe natural en contorno local, democratiza a explotación del indicador dentro da organización ao permitir o seu uso directo en informes estratéxicos sen requirir coñecementos técnicos de programación.

English abstract

Anticipating business cycle turning points is a critical priority for analysis and planning teams within the banking sector. However, the official measurement of aggregate activity through Gross Domestic Product (GDP) suffers from a significant publication lag. This lack of synchronization with real-time decision-making drives the development of this paper, which aims to design a high-frequency leading indicator focused on nowcasting the Spanish economy.

The proposed tool relies on a Dynamic Factor Model in state-space form, a methodology that reduces the high dimensionality of a large panel of monthly and quarterly indicators by extracting a single unobservable common factor. Prior to joint estimation using the Expectation-Maximization algorithm and the Kalman Filter, the raw series undergo seasonal adjustment, linearization, and variance stabilization.In addition, a variable screening process is conducted to identify the optimal panel, alongside various treatments addressing the high-volatility period caused by COVID-19.

Out-of-sample, pseudo-real-time validation demonstrates the indicator's consistency regarding historical inertia and forecasts from organizations using benchmark models, while also offering high economic interpretability through a marginal impact plot for each new data point; additionally, a sensitivity analysis is conducted to stress-test the model against low-probability scenarios. Finally, the model is integrated into an interactive web platform built with R Shiny. This dashboard, enhanced with reactive interfaces and a local natural language assistant, democratizes the indicator's deployment across the organization, allowing direct use in strategic reports without requiring technical programming expertise.

Capítulo 1

Introducción

1.1. Contexto y Motivación

El análisis de coyuntura y la anticipación de los puntos de giro del ciclo económico son esenciales en la gestión estratégica de las entidades financieras modernas. En el departamento de Planificación Estratégica y PMO de ABANCA, el seguimiento continuo del entorno macroeconómico resulta indispensable para la toma de decisiones, ya que determina las directrices de los planes estratégicos, los presupuestos anuales y el diseño de pruebas de estrés. Sin embargo, la medición de la actividad agregada enfrenta un obstáculo: el decalaje temporal de las estadísticas oficiales. Esta asincronía estructural genera una limitación operativa en el sector bancario, abriendo una brecha entre el momento en el que ocurren los choques económicos y su reflejo contable. Por ello, este proyecto plantea las siguientes cuestiones: ¿es posible desarrollar un indicador adelantado fiable de la economía española en tiempo real? ¿Qué variables y sectores aportan mayor información al crecimiento? ¿Bastan los modelos lineales para capturar los puntos de giro o es preciso transicionar hacia metodologías no lineales?

1.2. Metodología y objetivos del proyecto

El entorno macroeconómico está compuesto por una densa red de interrelaciones donde el mercado laboral, la producción industrial, las transacciones comerciales, las encuestas de opinión y los flujos financieros se mueven conjuntamente de forma simultánea. Realizar análisis individuales univariantes sobre cada indicador fragmenta la información y provoca una pérdida de las propiedades de conjunto del sistema de datos. Para resolver esta dispersión transversal la teoría estadística ofrece dos aproximaciones clásicas de reducción de la dimensionalidad: el Análisis de Componentes Principales (PCA) y el Análisis Factorial. Aunque ambos enfoques comparten la meta de comprimir la información de un panel extenso de dimensiones, a diferencia fundamental radica en que el PCA es una transformación orientada a maximizar la varianza total explicada sin discriminar la naturaleza del residuo, mientras que el Análisis Factorial constituye un modelo estadístico estructural diseñado específicamente para modelizar y explicar la estructura de las **covarianzas** y comovimientos entre las variables originales.

El Modelo Factorial Dinámico toma esta última filosofía y la generaliza hacia el dominio de las series temporales. El supuesto de partida asume que un panel de N indicadores macroeconómicos observables se encuentra conducido de forma latente por un número muy reducido de factores comunes inobservables ($r \ll N$). Para cada variable i en el instante temporal t, la observación se descompone de forma exacta mediante la suma lineal:

$$X_{it} = \lambda_{i1} f_{1t} + \lambda_{i2} f_{2t} + \cdots + \lambda_{ir} f_{rt} + e_{it}$$

Donde los coeficientes λ_{ij} representan las cargas factoriales que miden la sensibilidad de la variable frente al factor común f_{jt}, y la combinación lineal de los mismos constituye la **componente común**

de la serie. Por su parte, el término e_{it} recoge la perturbación **aleatoria**, que captura el ruido de medida y los movimientos particulares de esa variable específica. Al trasladar este diseño al Espacio de Estados, el factor inobservable se interpreta económicamente como el "estado real de la economía".

- **Objetivo Principal**: Desarrollar e implementar un indicador sintético de actividad en tiempo real (Nowcasting) enfocado en la economía española, explotando un panel de series macroeconómicas de alta frecuencia, con el fin de proporcionar a la entidad financiera una estimación temprana, robusta y automatizada frente al retraso de la estadística oficial del PIB.
- **Objetivos Específicos**:
  1. Diseñar un proceso que automatice la descarga y el preprocesamiento homogéneo de las series macroeconómicas sectoriales desde fuentes oficiales (INE, MINECO, Eurostat).
  2. Evaluar e implementar diferentes estrategias para estabilizar la volatilidad atípica causada por la pandemia de la COVID-19 en el año 2020.
  3. Comparar la precisión predictiva fuera de la muestra y la estabilidad paramétrica del MFD frente a las estimaciones oficiales de organismos de referencia (AIReF) y el benchmark univariante, un AR(1).
  4. Integrar el modelo final en una plataforma web interactiva corporativa (R Shiny) dotada de un asistente conversacional (LLM)[1] en entorno local, garantizando la seguridad de la información y democratizando la explotación analítica del indicador dentro de la organización.

1.2.1. Conceptos previos y terminología del Nowcasting

A continuación se definen los conceptos clave empleados en la modelización en tiempo real. Un panel de datos sectorial de coyuntura suele estar desbalanceado. A diferencia de las matrices teóricas cuadradas, las series de los organismos públicos no comparten el mismo inicio histórico. Además, el panel presenta bordes dentados (ragged edges) en su margen final debido a los distintos retrasos de publicación institucional. Dicha distorsión es el resultado directo de las asincronías inherentes a los calendarios institucionales de publicación. La estructura del panel descrita se muestra en la Figura 1.1. En este gráfico se aprecia el formato escalonado provocado por los diferentes retrasos de publicación de las variables, así como el porcentaje de valores ausentes de cada serie.

Las variables que integran el panel de indicadores de la entidad se clasifican según la naturaleza intrínseca del registro de obtención de los datos y su velocidad de difusión, en dos tipologías fundamentales:

- **Soft Data (Datos Cualitativos)**: Indicadores procedentes de encuestas de opinión, expectativas y sentimiento macroeconómico (tales como los índices de gestores de compras PMI o el Indicador de Confianza del Consumidor). Se caracterizan por publicarse de forma casi inmediata al cierre del mes de referencia y carecer de tendencia a largo plazo (tiene límites superiores e inferiores entre los valores que pueden tomar), actuando como las primeras alertas de los puntos de giro del ciclo.
- **Hard Data (Datos Cuantitativos)**: Registros oficiales de carácter administrativo o contable que miden variables reales de volumen, transacciones físicas o empleo efectivo (como el Índice de Producción Industrial, la Cifra de Negocios del Sector Servicios o las Afiliaciones a la Seguridad Social). Reflejan la masa contable real de la actividad productiva, pero sufren de mayores decalajes en su difusión oficial.

Por otro lado, se delimitan los conceptos que diferencian las tres perspectivas temporales de la predicción macroeconómica a muy corto plazo, representadas gráficamente en la Figura 1.2:

[1] Un Modelo de Lenguaje de Gran Escala (*Large Language Model*, LLM) es un modelo estadístico avanzado de Inteligencia Artificial basado en arquitecturas de aprendizaje profundo (redes tipo *Transformer*), entrenado con macroconjuntos de datos para modelizar la probabilidad de secuencias de palabras y generar lenguaje humano contextual.

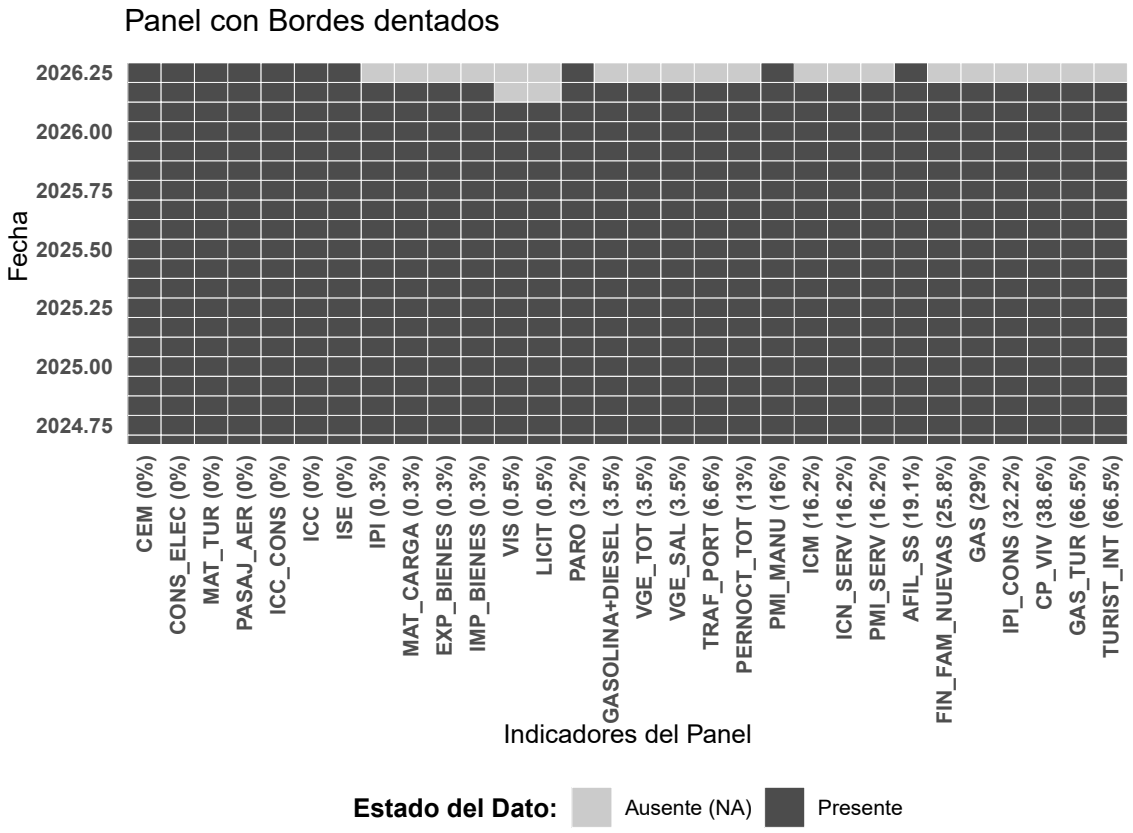


Figura 1.1: Panel desbalanceado de indicadores económicos.Elaboración propia.

.

- **Forecasting**: Se refiere a la predicción macroeconómica tradicional orientada hacia horizontes futuros $(t + h)$, proyectando la trayectoria del ciclo más allá del trimestre del que se están recibiendo datos concurrentes.
- **Nowcasting**: Es el concepto base de este proyecto. Se enfoca en la estimación e inferencia del valor actual del trimestre en curso (t), utilizando para ello la información parcial, asíncrona e intermitente mensual que va publicándose de forma progresiva a lo largo de esos tres meses.
- **Backcasting**: Consiste en la predicción o estimación del pasado reciente $(t-h)$. Este escenario se presenta cuando un trimestre ya ha finalizado cronológicamente pero el organismo oficial (INE) aún no ha difundido el dato contable del PIB, requiriendo rellenar dicho vacío con los últimos cierres de los indicadores mensuales sectoriales disponibles.

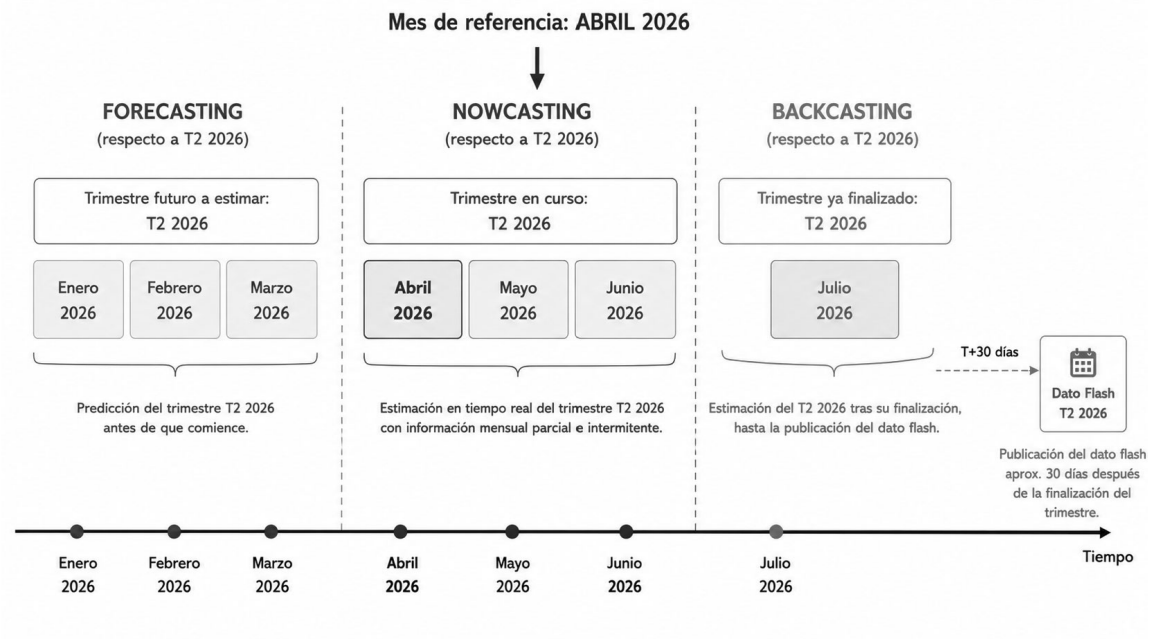


Figura 1.2: Identificación del Forecasting, Nowcasting y Backcasting dentro del trimestre de referencia. Elaboración propia.

.

Por último, la modelización debe considerar las revisiones de datos (vintages). Las estadísticas institucionales se actualizan a medida que se incorporan muestras más completas o cambios metodológicos. El nowcasting es muy sensible a estas revisiones, lo que obliga al analista a usar bases de datos de tiempo real para evitar sesgos retrospectivos en la evaluación fuera de la muestra.(Clements y Hendry, 2011).

1.3. Estructura de la Memoria

La presente memoria se estructura en dos grandes bloques metodológicos, distribuidos a lo largo de los siguientes capítulos:

La **Parte I** integra el marco teórico y los fundamentos del proyecto. El Capítulo 2 junto con el Apéndice A formaliza las propiedades de los procesos estocásticos univariantes y multivariantes clásicos, detallando la limitación paramétrica de esta modelización ante paneles de alta dimensión y presentando

el algoritmo X-13-ARIMA-SEATS para la extracción de la señal desestacionalizada, necesario para la entrada de datos en el modelo. El Capítulo 3 desarrolla en profundidad la formulación matemática del Modelo Factorial Dinámico en el Espacio de Estados, detallando la mecánica recursiva del Filtro de Kalman y los distintos métodos de estimación para resolver las diferentes situaciones en las que se encuentra uno en la aplicación real de este modelado.

La **Parte II** aborda la aplicación práctica sobre la economía española. El Capítulo 4 detalla la construcción del panel de variables, la automatización de las descargas y el preprocesamiento de las series (desestacionalización, estacionariedad y estandarización). El Capítulo 5 expone la selección y especificación del modelo, definiendo el tratamiento de los datos de la COVID-19 y la fijación de los parámetros estructurales.

A continuación, el Capítulo 6 evalúa el ajuste del modelo, su estabilidad paramétrica y la precisión predictiva intra-trimestral mediante la descomposición de noticias y pruebas de sensibilidad. Por último, el Capítulo 7 describe la integración tecnológica del modelo en una aplicación interactiva en R Shiny, incluyendo un asistente conversacional local, y recoge las conclusiones finales del proyecto.

Parte I

MARCO TEÓRICO Y METODOLOGÍA

Capítulo 2

Análisis de series temporales

En este capítulo se establecen los fundamentos teóricos y la estrategia que constituyen la base metodológica del presente proyecto. La caracterización formal de los datos económicos como series temporales es fundamental para modelizar su evolución dinámica y aislar el comovimiento subyacente (la señal común) del ruido aleatorio.

Para aligerar la lectura y centrar el capítulo en el desarrollo práctico del *nowcasting*, el rigor puramente matemático, las demostraciones y la formulación paramétrica estricta de los modelos univariantes y multivariantes han sido trasladados al **Apéndice A**, al cual se hará referencia de manera recurrente durante la presente memoria.

Para aligerar la lectura y centrar el capítulo en el desarrollo práctico del nowcasting, la formulación paramétrica estricta de los modelos univariantes y multivariantes se ha trasladado al Apéndice A. En este capítulo se introducen los conceptos fundamentales de series temporales, se analizan las limitaciones de dimensionalidad de los modelos multivariantes tradicionales (VAR) que justifican el uso de factores dinámicos, y se describe el algoritmo X-13ARIMA-SEATS empleado para la desestacionalización de los datos.

2.1. De la Modelización Univariante a la Multivariante

En este tipo de modelado los datos se registran en intervalos regulares de tiempo (mensual, trimestral). Dado que en economía solo observamos una única trayectoria de la historia (una muestra de tamaño $n = 1$ para cada instante t), la estadística exige asumir que el proceso se encuentra en un estado de equilibrio. Este equilibrio se formaliza mediante las propiedades de **estacionariedad** y **ergodicidad** (cuyas definiciones y teoremas se encuentran en el **Apéndice A**).

2.1.1. El enfoque univariante: Metodología Box-Jenkins

La modelización univariante asume que la mejor predicción del futuro de una variable reside exclusivamente en su propio pasado histórico y en los *shocks* (innovaciones) que ha sufrido. Apoyándose en el Teorema de Descomposición de Wold, la metodología clásica de Box, Jenkins, Reinsel, y Ljung (2015) propone capturar esta inercia mediante los siguientes modelos:

- **Modelos Autorregresivos (AR):** El valor actual de la serie se explica como una combinación lineal de sus propios valores pasados.
- **Modelos de Medias Móviles (MA):** La observación actual responde a la innovación (ruido blanco) presente y a los *shocks* pasados.
- **Modelos ARIMA Estacionales:** La síntesis de las estructuras anteriores que, tras aplicar operadores de diferenciación regular y estacional para estabilizar la media, permite ajustar dinámicas

complejas que combinan memoria a corto plazo y memoria estacional (véase formulación exacta en el **Apéndice A**).

Sin embargo, para el ejercicio del *nowcasting*, el enfoque univariante adolece de una "ceguera" estructural. Dada la baja frecuencia de publicación del PIB (trimestral y con un rezago cercano a los 30 días), a un modelo ARIMA del PIB le resulta complicado reaccionar ante *shocks* contemporáneos, perdiendo la capacidad de alerta temprana que se requiere en algunas instituciones para realizar previsiones.

2.1.2. El enfoque multivariante: El Modelo VAR

La evolución lógica conduce hacia la modelización multivariante, donde las magnitudes macroeconómicas no evolucionan de forma aislada, sino impulsadas por choques estructurales compartidos. El modelo de **Vectores Autorregresivos (VAR)** captura esta interdependencia tratando a todas las variables de un vector $\mathbf{Y_t}$ como endógenas y explicándolas en función de los retardos de todas las variables del sistema (Sims, 1980) (véase la formulación exacta en el **Apéndice A**).

Observación 2.1. *Para facilitar la lectura de esta sección, se ha omitido la descripción detallada de los operadores de álgebra lineal y la notación matricial en este cuerpo de texto. Se remite al lector al Apéndice B para una exposición completa de la notación y operadores matemáticos empleados tanto en el Apéndice explicativo del modelo multivariante como en la metodología de factores dinámicos posterior.*

A pesar de la solidez teórica de los modelos multivariantes tradicionales para sistemas de pequeña escala, su aplicación en la predicción de una variable de interés como el Producto Interior Bruto (PIB) mediante paneles de alta dimensión encuentra un límite conocido como la *maldición de la dimensionalidad*. Este fenómeno, acuñado originalmente por (Bellman, 1961), describe cómo la complejidad y la dispersión de los datos crecen de forma exponencial al añadir nuevas variables, degradando la capacidad predictiva y la significatividad estadística del modelo.

2.1.3. Definición Formal y Explosión Paramétrica

En el marco de la macroeconometría, esta limitación se manifiesta a través de la explosión del espacio de parámetros. Para un modelo VAR(p) con N variables endógenas, el número total de parámetros a estimar (K) se deriva de la suma de los interceptos de cada ecuación y las matrices de coeficientes de cada retardo:

$$K = N + pN^2,$$

esta expresión revela un crecimiento cuadrático respecto al número de variables (N). La justificación matemática reside en la estructura de interdependencia del modelo: cada nueva variable añadida al sistema no solo aporta su propia ecuación, sino que exige cuantificar su interacción con todos los retardos de todas las variables ya existentes. Mientras que el coste de aumentar los retardos (p) es lineal, el coste de aumentar la dimensión informativa (N) es exponencialmente más gravoso para la eficiencia del estimador.

2.1.4. Dispersión en el Espacio y Densidad Muestral (*Sparsity)*

El impacto más crítico de la alta dimensionalidad es la dispersión geométrica de las observaciones. A medida que aumenta el número de variables (N), el volumen del espacio paramétrico crece exponencialmente, provocando que la densidad de los datos muestrales disponibles disminuya de forma drástica. En términos estadísticos, esta dispersión (sparsity) implica que el modelo carece de observaciones suficientes para estimar relaciones fiables. El sistema intenta ajustar miles de coeficientes en un espacio con escasa información efectiva, lo que se traduce en una pérdida crítica de grados de libertad, inflando la varianza de los estimadores y generando problemas severos de sobreajuste

2.1.5. Análisis Comparativo y Agotamiento de Grados de Libertad

Para ilustrar la problemática que sufren los modelos multivariantes tradicionales frente a la escala necesaria para capturar la actividad económica agregada de España, comparamos dos escenarios:

- Modelo de Referencia (Escala Reducida): El enfoque de (Blanco Bugueiro, 2024) para el IPC utilizó una desagregación de 5 componentes fundamentales. Al ajustar un modelo VAR(4), el sistema requirió solo 105 coeficientes ($5 + 4 \times 5^2$). Con una muestra de aproximadamente 200 observaciones, la densidad de datos es suficiente para garantizar la consistencia.
- Este Proyecto (Escala Ampliada): Para realizar un *nowcasting* preciso de la actividad económica española puede ser necesario monitorear un panel de un número relativamente elevado de series, pongamos por ejemplo un número razonable de 40 indicadores económicos (producción industrial, consumo eléctrico, confianza del consumidor, ventas minoristas, etc.)–en otros trabajos similares trabajan con cientos de indicadores–. Bajo una especificación similar de 4 retardos, el modelo exigiría:

  $$40 + 4(40^2) = \mathbf{6{,}440} \text{ coeficientes}$$

  Incluso disponiendo de 20 años de historia mensual ($T = 240$), el número de parámetros a estimar superaría en casi 27 veces el número de observaciones disponibles, por lo tanto $T \ll K$. En este escenario, el modelo sufriría un sobreajuste (overfitting) total, memorizando el ruido de la muestra en lugar de aprender la señal económica de fondo. Otro ejemplo real de este tipo de modelado cuando se pretende reducir a un indicador el quehacer de una sistema tan completo y enorme como la economía de EEUU, tal como detallan Stock y Watson (1989), «*The leading variables were chosen from an initial list of approximately 280 series (Mitchell and Burns (1938) started with 487 series)*», la explosión paramétrica que resultaría de un modelo tipo VAR con esta cantidad de parámetros a estimar sería computacionalmente muy demandante, aunque se redujeron finalmente a la mitad el número de variables y de parámetros a estimar la situación sería tremendamente problemática.

Esta restricción paramétrica supone una pérdida de información relevante (y su consecuente sesgo en las estimaciones)(Bernanke, Boivin, y Eliasz, 2005). Dado que los modelos VAR estándar no pueden integrar paneles macroeconómicos de gran escala sin riesgo de sobreajuste o problemas de invertibilidad (no fundamentalidad)[1], los Modelos Factoriales Dinámicos (MFD) ofrecen una alternativa robusta. En lugar de estimar las N^2 interacciones individuales, los MFD asumen que la dinámica de los indicadores está impulsada por un número reducido de factores latentes ($r \ll N$), recuperando la parsimonia y la densidad muestral sin sacrificar la información del panel original (Stock y Watson, 2002a).

2.2. Algoritmo de descomposición de series temporales:X-13-ARIMA-SEATS

En el análisis macroeconómico, se asume que una serie temporal $\{X_t\}$ es el resultado de la interacción de diversos factores no observables que operan en diferentes horizontes temporales. La descomposición de series temporales en factores no observables tiene su origen formal en los trabajos de (Persons, 1919), quien estableció el esquema de cuatro componentes: tendencia (T_t), ciclo (C_t), estacionalidad (S_t) e irregularidad (I_t). Esta estructura clásica permite capturar la señal de la serie temporal limpia de ruidos estacionales y eventos fortuitos. Las componentes se definen como sigue:

- **Tendencia (T_t):** Representa la evolución o dirección de la serie a largo plazo, reflejando el crecimiento o decrecimiento sostenido de la economía.

[1] El problema de invertibilidad surge cuando la información contenida en un número reducido de variables no es suficiente para recuperar los choques estructurales de la economía. Como señalan (Bernanke et al., 2005) y (Stock y Watson, 2002b), el uso de Modelos Factoriales permite mitigar este problema al expandir el conjunto de información explotada, asegurando que el espacio generado por los factores abarque el de los verdaderos componentes estructurales.

- **Ciclo (C_t):** Recoge las oscilaciones de medio plazo alrededor de la tendencia, vinculadas a las fases de expansión y recesión económica. En la práctica, tendencia y ciclo suelen tratarse conjuntamente como la componente *ciclo-tendencia*.
- **Estacionalidad (S_t):** Engloba los comportamientos que se repiten de forma regular en periodos iguales o inferiores a un año (por ejemplo, el aumento del consumo en navidad o el efecto de las vacaciones de verano).
- **Irregularidad (e_t):** También denominada componente aleatoria, alberga los efectos residuales resultantes de hechos no previsibles o errores de medición.

La combinación de estas componentes se formaliza mediante dos esquemas principales (se ha optado por juntar Tendencia y Ciclo en una sola componente como se hace en la práctica habitual):

1. **Modelo Aditivo:** Se utiliza cuando la magnitud de las oscilaciones estacionales es constante e independiente del nivel de la serie:
$$X_t = T_t + S_t + e_t.$$
2. **Modelo Multiplicativo:** Es el esquema más frecuente en series económicas, donde la variabilidad estacional aumenta o disminuye de forma proporcional al nivel (tendencia) de la serie[2]:
$$X_t = T_t \cdot S_t \cdot e_t.$$

2.2.1. El estándar institucional: X-13-ARIMA-SEATS

Para llevar a cabo esta descomposición, es decir, para extraer los componentes inobservables $T_t, S_t y e_t$, se emplea el algoritmo **X-13ARIMA-SEATS** (abreviado habitualmente como X-13) (U.S. Census Bureau, 2017). Este software integra las dos vertientes de desestacionalización[3]: la aproximación de medias móviles desarrollada por la Oficina del Censo de los Estados Unidos (*US Census Bureau*) a través de sus históricos filtros X-11 y X-12-ARIMA, y la escuela de modelización paramétrica desarrollada por el Banco de España mediante el programa TRAMO-SEATS. En la actualidad, constituye el algoritmo de desestacionalización más versátil y con mayor renombre internacional, siendo utilizado por los principales organismos estadísticos, incluyendo Eurostat (Eurostat, 2015) y el Instituto Nacional de Estadística español (Instituto Nacional de Estadística, 2019).

Lo más destacable de la metodología es la gran flexibilidad que proporciona al investigador:

2.2.1.1. Fase 1: Preajuste y Linearización (Módulo regARIMA)

Los filtros matemáticos encargados de separar la estacionalidad son extremadamente sensibles a los valores atípicos y a la varianza no constante. Por ello, este módulo inicial ajusta preliminarmente un modelo de regresión con variables regresoras deterministas y errores estocásticos (conocido como modelo *regARIMA*) para "linearizar" la serie. Esta integración tiene su origen en los trabajos seminales de Gómez y Maravall (1996) para el programa TRAMO, posteriormente expandidos por el *US Census Bureau* (Findley, Monsell, Bell, Otto, y Chen, 1998).

Matemáticamente, el modelo regARIMA asume que la serie temporal observada (o una transformación logarítmica de la misma), denotada como y_t, puede descomponerse en una combinación lineal de efectos deterministas y un término de error estocástico z_t:

[2]De forma natural se observa como el modelo multiplicativo aplicando logaritmos a sus componentes se convierte en el modelo aditivo

[3]La desestacionalización es el proceso estadístico mediante el cual se aísla y elimina el componente estacional (S_t) de una serie temporal. Este componente agrupa las fluctuaciones periódicas y predecibles que se repiten con un patrón anual (por ejemplo, el aumento del consumo en Navidad o los repuntes de contratación en la campaña estival). Su eliminación permite aislar la tendencia subyacente y el ciclo económico real, posibilitando una comparación homogénea de los datos entre periodos consecutivos.

$$y_t = \sum_{i=1}^{k} \beta_i x_{it} + z_t,$$

donde x_{it} representa un conjunto de k regresores exógenos (que modelan los efectos de calendario y los valores atípicos), β_i son los coeficientes de regresión asociados, y z_t es el componente de ruido estocástico. A diferencia de una regresión lineal clásica por Mínimos Cuadrados Ordinarios (MCO), el término de error z_t no se asume ruido blanco, sino que sigue un proceso estocástico estacional $SARIMA(p,d,q) \times (P,D,Q)_s$:

$$\phi_p(B)\Phi_P(B^s)(1-B)^d(1-B^s)^D z_t = \theta_q(B)\Theta_Q(B^s)a_t,$$

donde B es el operador de retardo ($B^k z_t = z_{t-k}$); s es la frecuencia estacional de la serie ($s = 12$ para datos mensuales); $\phi_p(B)$ y $\Phi_P(B^s)$ son los polinomios autorregresivos regular y estacional; $\theta_q(B)$ y $\Theta_Q(B^s)$ son los polinomios de medias móviles; $(1-B)^d$ y $(1-B^s)^D$ son los operadores de diferencias necesarios para inducir estacionariedad; y a_t es un proceso de ruido blanco con media nula y varianza σ_a^2.

A través de esta formulación, el algoritmo identifica en fecha y estima (los valores β_i) el impacto de los efectos de calendario, al tiempo que modeliza de forma automática las alteraciones estructurales y los datos atípicos. Estas perturbaciones exógenas se clasifican en tres tipologías fundamentales:

- **Atípicos Aditivos (AO):** Un shock estrictamente puntual en el instante t.
- **Cambios de Nivel (LS):** Una alteración estructural que provoca un desplazamiento permanente en la media local a partir de un instante conocido.
- **Cambios Transitorios (TC):** Perturbaciones con una decadencia exponencial que se disipa asintóticamente.

El propósito de esta Fase 1 es obtener una serie linealizada libre de todo efecto determinista y atípico, garantizando que el motor de extracción de la Fase 2 estime la componente estacional pura (S_t) sin distorsiones ni sesgos provocados por quiebres estructurales o atípicos. Una vez aislado el factor estacional, el algoritmo calcula la **Serie Corregida de Variaciones Estacionales y de Calendario (CVEC)** sustrayendo exclusivamente la estacionalidad y el calendario de la serie original. En consecuencia, y por definición metodológica, las perturbaciones de carácter estructural y coyuntural (`AO`, `LS` y `TC`) permanecen integradas en la serie final CVEC para preservar los impactos históricos reales sobre las magnitudes económicas.

Adicionalmente, el proceso estocástico estimado para z_t permite generar predicciones óptimas hacia el futuro (*forecasts*) y hacia el pasado (*backcasts*) de la serie. Este paso es crítico para mitigar la asimetría y el sesgo de revisión de los filtros de medias móviles en los extremos de la muestra durante la posterior extracción de la señal.

2.2.1.2. Fase 2: Extracción de la Señal (Motores SEATS y X-11)

Una vez se obtiene la serie z_t, el algoritmo transita hacia la fase de descomposición. La principal innovación del X-13 es que permite al investigador seleccionar el motor de extracción de señal más adecuado según la naturaleza del ruido.

Por un lado, puede emplear el motor **SEATS**, basado puramente en la extracción paramétrica de señales a partir del modelo ARIMA ajustado en la fase anterior. Por otro lado, frente a series que exhiben varianzas altas o ruido extremo, el algoritmo permite recurrir a la flexibilidad del filtrado no paramétrico mediante el motor **X-11**[4].

[4]Esta metodología sigue una secuencia iterativa de filtros de medias móviles centradas para descomponer la serie en tendencia-ciclo, estacional e irregular. Su robustez frente a especificaciones erróneas lo convierte en una alternativa invaluable cuando los modelos paramétricos fallan en su convergencia.

En su variante X-11, el algoritmo aplica iterativamente una secuencia de filtros simétricos. En concreto, emplea los filtros de Henderson[5] para extraer la curva de la tendencia-ciclo (TC_t), y filtros estacionales específicos (como el filtro 3×5 o variantes más agresivas como el 3×9)[6] para aislar los factores estacionales (S_t). El resultado final de este proceso es la obtención de la **Serie Corregida de Variaciones Estacionales y de Calendario (CVEC)**.

[5]Los filtros de Henderson son un tipo particular de media móvil simétrica diseñada matemáticamente para extraer la componente ciclo-tendencia. Su principal virtud frente a una media móvil simple es que consiguen suavizar el ruido irregular sin distorsionar la trayectoria parabólica local de la tendencia ni introducir rezagos de fase.

[6]La nomenclatura $n \times m$ denota una media móvil de n términos aplicada sobre una media móvil previa de m términos, calculadas para cada mes de forma independiente a lo largo de los años (por ejemplo, promediando todos los eneros). El filtro 3×5 es el estándar y asume una estacionalidad que cambia gradualmente. Por su parte, el filtro 3×9 abarca una ventana temporal más larga, generando un factor estacional mucho más suave y estable. El uso de filtros amplios es preceptivo en series con una varianza irregular extrema (como las disrupciones turísticas por la pandemia), ya que evita que el algoritmo confunda el ruido transitorio con un verdadero cambio estructural en el patrón estacional.

Capítulo 3

El Modelo Factorial Dinámico

La elección del MFD como pilar metodológico de este proyecto responde a la necesidad de superar los obstáculos inherentes al análisis macroeconómico de alta dimensión en la predicción a muy corto plazo. Emplear un modelo de regresión lineal múltiple tradicional, estimado mediante mínimos cuadrados ordinarios, para proyectar el PIB utilizando directamente todo el panel de indicadores resulta inviable cuando el número de variables explicativas (N) es elevado. En este escenario, el modelo agota rápidamente los grados de libertad y sufre un severo problema de sobreajuste provocado por la multicolinealidad. Como consecuencia, el error cuadrático medio de predicción fuera de la muestra deja de reducirse, lo que refleja un desequilibrio crítico entre el sesgo y la varianza. El MFD resuelve matemáticamente esta "maldición de la dimensionalidad" —como se justificó en la Sección 2.1.4— al condensar la vasta información del panel en un número muy reducido de factores latentes inobservables ($r \ll N$), los cuales logran capturar la dinámica cíclica común del sistema.

Una de las virtudes más destacadas de este enfoque es su capacidad para gestionar las irregularidades propias de los datos en tiempo real. Frente a los modelos clásicos que exigen paneles de datos balanceados, el MFD, en su forma de Espacio de Estados, permite trabajar con bordes dentados (*ragged edges*) -fenómeno derivado de los calendarios asíncronos de publicación de las series temporales económicas- y con frecuencias mixtas, integrando indicadores mensuales en la predicción de variables trimestrales como el PIB a través de la transformación de Mariano y Murasawa (2003). Mediante el uso del Filtro de Kalman, el modelo actúa como un estimador óptimo de mínima varianza, actualizando recursivamente la señal económica conforme se publica cada nuevo dato.

La aplicación práctica del MFD ha sido probada satisfactoriamente por instituciones de referencia como el Banco de España a través del conocido modelo *Spain-STING*(Arencibia Pareja, Gómez Loscos, De Luis López, y Pérez-Quirós, 2024; Camacho y Pérez-Quirós, 2011) y la Reserva Federal de EE.UU. con su *New York Fed Staff Nowcast* iniciado en el año 2016 (Aarons et al., 2016).

A continuación, la estructura de este capítulo transita desde la introducción conceptual y matemática del modelado en el Espacio de Estados, ya que parece adecuado presentar esta metodología tan amplia de forma individualizada y algo más extensa, para luego ir hacia la operatividad del modelo desde distintas perspectivas. En primer lugar, en la Sección 3.1 se define la representación formal de los sistemas lineales en el Espacio de Estados y su evolución histórica desde la ingeniería de control hasta la econometría clásica de Kalman (1960), Harvey (1989) y una referencia más cercana en el tiempo de Durbin y Koopman (2012), terminando con el algoritmo recursivo de inferencia dentro de esta representación, el filtro de Kalman. Seguidamente, en la Sección 3.3 se presenta la representación del MFD en sus distintas formas y varios de los métodos de estimación, tanto paramétrico como no paramétrico. El capítulo sigue detallando varios criterios para determinar el número de factores latentes (Bai y Ng, 2002, 2007), así como el tratamiento de las frecuencias mixtas necesario para el *nowcasting* del PIB español. Asimismo, cabe destacar la influencia de los trabajos fundamentales de Stock y Watson (1989, 2002b, 2011, 2016), y uno los exponentes más importantes en la actualidad de estos modelos y antiguo empleado de la Reserva Federal de EE.UU. y del Banco Central Europeo (BCE), Domenico

Giannone,(Giannone, Reichlin, y Small, 2008).

3.1. Introducción al modelado en el Espacio de Estados

En la presente sección se introducirá a la representación de modelos en el Espacio de Estados, la cual se trata de un sistema de representación de modelos tremendamente completo, por tanto, se cree que merece una introducción propia y más profunda.

3.1.1. Evolución histórica: de Wiener a Harvey, pasando por Kalman

El origen de esta metodología se sitúa en el campo de la ingeniería de control y la industria aeroespacial. Antes del trabajo fundamental de Kalman en 1960, el paradigma dominante era el enfoque de Norbert Wiener en la década de 1940 (Wiener, 1949). El método de Wiener se basaba en la factorización espectral en el dominio de la frecuencia, pero presentaba limitaciones críticas: requería que las estadísticas fueran estrictamente estacionarias y su implementación resultaba computacionalmente compleja para problemas de gran escala. El cambio de paradigma llegó con el artículo de Kalman (1960). Kalman reexaminó el problema del filtrado introduciendo el método de *transición de estados*, trasladando el análisis al dominio del tiempo. A diferencia del filtro de Wiener, el filtro de Kalman es un algoritmo recursivo capaz de procesar tanto magnitudes estacionarias como no estacionarias, lo que significó un salto cualitativo para la aplicación práctica en sistemas dinámicos. A partir de la década de 1980, el interés por esta metodología creció notablemente en la econometría. La figura clave fue Andrew Harvey, quien a través de sus obras (Harvey, 1981, 1989) introdujo formalmente el uso del filtro de Kalman para obtener estimaciones de máxima verosimilitud mediante la descomposición del error de predicción. Harvey demostró que una vasta gama de modelos econométricos -incluyendo los procesos ARMA, modelos de componentes no observados y modelos multivariantes- podían expresarse en forma de Espacio de Estados, convirtiéndola en una herramienta estándar para la extracción de señales económicas inobservables.

3.1.2. Fundamentos de la representación de modelos en el Espacio de Estados (SS)

La modelización de sistemas dinámicos mediante la formulación de Espacio de Estados constituye un marco unificado para el tratamiento de un amplio rango de problemas en el análisis de series temporales. En este enfoque, se asume que el desarrollo temporal de un sistema bajo estudio está determinado por una serie de vectores inobservables denominados *estados* ($\boldsymbol{\alpha}_t$).

Definición 3.1. *Un estado, $\boldsymbol{\alpha}_t$, se define como el subconjunto mínimo de variables $\{x_1(t), x_2(t), \ldots, x_n(t)\}$ tales que su conocimiento en un instante t_0, junto con la señal de entrada para $t \geq t_0$, es suficiente para determinar el comportamiento del sistema en cualquier tiempo futuro.*

Estos estados se vinculan con las observaciones reales (y_t) a través de un sistema de ecuaciones lineales. A diferencia de los métodos econométricos tradicionales que se centran en las variables observables, la formulación en el Espacio de Estados permite modelar componentes latentes -como tendencias, ciclos económicos o factores comunes- que evolucionan siguiendo una lógica de los procesos de Markov de primer orden (Markov, 1906), que se definen tal como:

Definición 3.2. *Un proceso estocástico $\{\boldsymbol{\alpha}_t\}$ posee la propiedad de Markov de primer orden si la distribución de probabilidad del estado futuro, condicionado tanto a los estados pasados como al presente, depende exclusivamente de la situación del sistema en el instante actual. Formalmente, se define mediante la igualdad de las densidades de probabilidad condicionales:*

$$P(\boldsymbol{\alpha}_{t+1}|\boldsymbol{\alpha}_t, \boldsymbol{\alpha}_{t-1}, \ldots, \boldsymbol{\alpha}_1) = P(\boldsymbol{\alpha}_{t+1}|\boldsymbol{\alpha}_t)$$

Esta naturaleza markoviana permite que el vector de estado $\boldsymbol{\alpha}_t$ funcione como un estadístico suficiente de la historia del sistema, condensando toda la información histórica relevante para determinar su evolución futura.

3.1.3. Estructura del Sistema de dos ecuaciones

Siguiendo la formulación de Durbin y Koopman (2012), un **Sistema Lineal Gaussiano en el Espacio de Estados** se define mediante dos ecuaciones vectoriales fundamentales que operan en tiempo discreto y que servirá como base para la representación posterior del MFD:

- **Ecuación de Medida (u Observación):** Describe la relación lineal entre el vector de observaciones y el estado latente:

$$\boldsymbol{y_t} = \boldsymbol{Z_t}\boldsymbol{\alpha}_t + \boldsymbol{\epsilon}_t, \quad \boldsymbol{\epsilon}_t \sim N(0, \boldsymbol{H_t}),$$

  donde $\boldsymbol{y}_t$ es un vector de observaciones de dimensión $p\times 1$, $\boldsymbol{\alpha}_t$ es un vector de estado no observable de dimensión $m \times 1$, $\boldsymbol{Z}_t$ es la matriz de medida $(p \times m)$ y $\boldsymbol{\epsilon}_t$ es el vector de perturbaciones irregulares o error de observación, asumido serialmente independiente.

- **Ecuación de Transición (o de Estado):** Determina la evolución dinámica del sistema mediante un proceso de Markov de primer orden:

$$\boldsymbol{\alpha}_{t+1} = \boldsymbol{T}_t\boldsymbol{\alpha}_t + \boldsymbol{R}_t\boldsymbol{\eta}_t, \quad \boldsymbol{\eta}_t \sim N(0, \boldsymbol{Q_t}),$$

  donde $\boldsymbol{T}_t$ es la matriz de transición $(m \times m)$, $\boldsymbol{\eta}_t$ es el vector de perturbaciones del estado $(r \times 1)$ y $\boldsymbol{R}_t$ es una matriz de selección $(m \times r)$ que vincula los choques estocásticos con las variables de estado correspondientes.

Tal como señala Harvey (1989), al conjunto de matrices que gobiernan estas ecuaciones ($\boldsymbol{Z_t}, \boldsymbol{T_t}, \boldsymbol{R_t}, \boldsymbol{H_t}, \boldsymbol{Q_t}$) se las denomina **matrices del sistema**. A menos que se especifique lo contrario, se asume que no son estocásticas (son deterministas), aunque pueden variar con el tiempo. Si estas matrices permanecen constantes para todo t, el modelo se clasifica como **invariante en el tiempo**, que será la clase de modelo seleccionado en la aplicación empírica de este proyecto.

Esta estructura permite inferir las propiedades del estado del sistema a partir de magnitudes observables contaminadas por componentes ruidosos. La representación a través de un diagrama de flujos de este sistema de ecuaciones en tiempo discreto con parámetros invariantes en el tiempo, la presentó Kalman en su famoso artículo científico de 1960. Esta representación se observa en la Figura 3.1, en la cual se ilustra el flujo de información desde el estado pasado $\boldsymbol{\alpha}_{t-1}$ hacia el presente y su vinculación con la observación y_t.

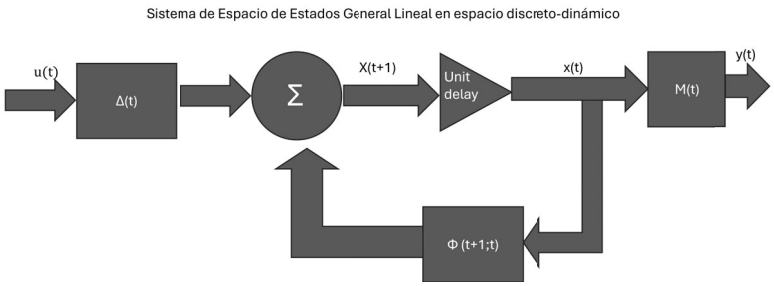


Figura 3.1: Representación del sistema dinámico en el Espacio de Estados según Kalman (1960).

.

Además, siguiendo con el enfoque de Durbin y Koopman (2012), una forma adecuada de introducir las aplicaciones directas en el espacio de estados es a través del Modelo de Nivel Local. Este modelo representa el caso más simple de un modelo estructural de series temporales expresadas en el Espacio de Estados, y sirve como ejemplo ilustrativo básico para una de las descomposiciones de series temporales univariantes como las descritas en la Sección 2.2. En esta especificación, se asume que la observación se compone únicamente de una tendencia estocástica y un error irregular, sin presencia de estacionalidad, ciclos o pendiente determinista. El estado α_t es un escalar que representa el nivel actual de la serie y evoluciona siguiendo un paseo aleatorio.

El sistema se define mediante las siguientes ecuaciones:

- **Ecuación de medida:**
$$y_t = \alpha_t + \epsilon_t, \quad \epsilon_t \sim N(0, \sigma_\epsilon^2).$$
- **Ecuación de transición:**
$$\alpha_{t+1} = \alpha_t + \eta_t, \quad \eta_t \sim N(0, \sigma_\eta^2).$$

Donde las perturbaciones ϵ_t y η_t son mutuamente independientes entre sí y con respecto al estado inicial α_1.

A pesar de su aparente sencillez, este modelo ilustra características fundamentales del espacio de estados:

1. **Invarianza temporal y No estacionariedad:** Es importante destacar, siguiendo a Harvey (1989), que aunque este es un modelo invariante en el tiempo (sus matrices del sistema son constantes: $Z_t = 1, T_t = 1$), genera un proceso **no estacionario** debido a la naturaleza de paseo aleatorio del estado. La varianza de α_t crece con el tiempo, capturando la evolución de las series económicas, aunque su forma estacionaria puede recuperarse mediante diferenciación.
2. **Extracción de señal:** Permite que el Filtro de Kalman extraiga el componente latente de forma óptima. Un concepto clave introducido por Durbin y Koopman (2012) en este contexto es la **ratio señal-ruido** ($q = \sigma_\eta^2/\sigma_\epsilon^2$). Este parámetro determina la sensibilidad del filtro: un q bajo implica una tendencia muy suave (el filtro confía poco en las nuevas innovaciones), mientras que un q alto permite que el nivel se ajuste rápidamente a los cambios recientes en los datos.

Este modelo sirve como base conceptual para entender cómo el espacio de estados puede descomponer series complejas en componentes inobservables, estableciendo el puente hacia una de las posibles representaciones y estimaciones de los Modelos Factoriales Dinámicos donde el "nivel" será compartido por múltiples series.

3.2. El filtro de Kalman

Una vez definido el sistema dinámico en el Espacio de Estados en la sección anterior, el problema central se reduce a la estimación del vector de estado inobservable $\boldsymbol{\alpha}_t$ condicionado al conjunto de información disponible $\boldsymbol{Y}_t = \{y_1, \ldots, y_t\}$. La solución óptima a este problema en su representación lineal en el Espacio de Estados, que supera las limitaciones de los filtros de Wiener clásicos al no requerir estacionariedad ni series infinitas, es el Filtro de Kalman. Introducido en el artículo seminal de Kalman (1960), nombrado ya previamente, este algoritmo no es simplemente una herramienta de cálculo, sino una derivación teórica basada en la proyección ortogonal en espacios de Hilbert[1]. A continuación, se detallan sus fundamentos teóricos, su mecánica recursiva y sus propiedades estadísticas.

[1]Formalmente, un espacio de Hilbert $\mathcal{H}$ es un espacio vectorial real o complejo dotado de un producto interior $\langle \cdot, \cdot \rangle$ que induce una norma $\|x\| = \sqrt{\langle x, x \rangle}$, respecto a la cual el espacio es métricamente completo (es decir, toda sucesión de Cauchy en $\mathcal{H}$ converge a un límite en $\mathcal{H}$). En el contexto de la teoría de filtrado y predicción lineal (Kalman, 1960), el espacio de interés es el espacio $L^2(\Omega, \mathcal{F}, P)$ formado por variables aleatorias de segundo orden (con varianza finita). En este marco, el producto interior se define como la esperanza matemática del producto cruzado, $\langle X, Y \rangle = \mathbb{E}[XY]$, lo que permite generalizar la geometría euclidiana a dimensiones infinitas. La propiedad de completitud garantiza que, dado un subespacio cerrado $\mathcal{Y}_t$ generado por las observaciones, existe un único elemento $\hat{\alpha}_t \in \mathcal{Y}_t$ (la proyección ortogonal) que minimiza la norma del error $\|\alpha_t - \hat{\alpha}_t\|$.

3.2.1. Fundamentos Teóricos: Proyección Ortogonal y Optimalidad

Para comprender la naturaleza del estimador, es necesario remitirse a los teoremas fundamentales presentados por Kalman. Sea $\mathcal{Y}_t$ la variedad lineal [2] generada por las observaciones $\{y_1, \ldots, y_t\}$. El problema de estimación consiste en encontrar un vector $a_{t|t}$ que minimice una función de pérdida escalar $L(\epsilon)$, donde $\epsilon = \alpha_t - a_{t|t}$, es el error de estimación.

Kalman (1960) establece en su Teorema 2 que la estimación óptima es la proyección ortogonal del estado sobre el espacio de las observaciones.

Teorema 3.1 (Teorema 2 de Kalman 1960)**.** *Si los procesos estocásticos $\{\alpha_t\}$ e $\{y_t\}$ son Gaussianos, o si la estimación se restringe a funciones lineales minimizando el error cuadrático medio, entonces el estimador óptimo $a_{t|t}$ es la proyección ortogonal de α_t sobre $\mathcal{Y}_t$:*

$$a_{t|t} = \mathbb{E}[\alpha_t | Y_t] = Proy(\alpha_t | \mathcal{Y}_t).$$

Esto implica que el error de estimación es ortogonal (incorrelado) a cualquier estimador basado en Y_t.

De este teorema y del Teorema 1 del mismo autor se derivan las propiedades de optimalidad de este algoritmo:

1. **Bajo Gaussianidad:** Si las perturbaciones ϵ_t y η_t son normales, el filtro proporciona el estimador de Mínimo Error Cuadrático Medio (MMSE), coincidiendo con la esperanza matemática condicional.

2. **MVLUE (Minimum Variance Linear Unbiased Estimator):** Aun relajando el supuesto de normalidad (como sugieren (Durbin y Koopman, 2012)), si el modelo mantiene la linealidad y los momentos finitos, el Filtro de Kalman sigue siendo el mejor estimador lineal posible (de mínima varianza) en este caso concreto.

Para que la proyección ortogonal descrita en el Teorema 2 garantice la optimalidad del estimador, (Kalman, 1960) exige la especificación rigurosa de las propiedades estocásticas del sistema.

Se asumen las siguientes condiciones sobre el modelo de Espacio de Estados:

1. **Distribución del Estado Inicial:** El algoritmo requiere un punto de partida probabilístico. Se asume que el vector de estado latente en el instante inicial, $\boldsymbol{\alpha}_1$, es un vector aleatorio con media incondicional conocida a_1 y matriz de varianzas-covarianzas $\boldsymbol{P}_1$, tal que $\mathbb{E}[\boldsymbol{\alpha}_1] = a_1$ y $Var(\boldsymbol{\alpha}_1) = \boldsymbol{P}_1$.

2. **Perturbaciones de Ruido Blanco (Ortogonalidad Temporal):** Los vectores de innovaciones ($\boldsymbol{\eta}_t$) y perturbaciones aleatorias ($\boldsymbol{\epsilon}_t$) constituyen procesos estocásticos de ruido blanco de media cero. Carecen de autocorrelación temporal:

   $$\mathbb{E}[\boldsymbol{\eta}_t \boldsymbol{\eta}_s'] = \mathbf{0} \quad \text{y} \quad \mathbb{E}[\boldsymbol{\epsilon}_t \boldsymbol{\epsilon}_s'] = \mathbf{0} \quad \forall t \neq s.$$

   Con matrices de covarianza contemporánea $\mathbb{E}[\boldsymbol{\eta}_t \boldsymbol{\eta}_t'] = \mathbf{H}_t$ y $\mathbb{E}[\boldsymbol{\epsilon}_t \boldsymbol{\epsilon}_t'] = \mathbf{Q}_t$.[3]

3. **Independencia Mutua Estricta:** No existe retroalimentación cruzada. Las perturbaciones del sistema, las de medida y el estado inicial son mutuamente incorrelados para todo t, s:

   $$\mathbb{E}[\boldsymbol{\eta}_t \boldsymbol{\epsilon}_s'] = \mathbf{0}, \quad \mathbb{E}[\boldsymbol{\eta}_t \boldsymbol{\alpha}_1'] = \mathbf{0}, \quad \mathbb{E}[\boldsymbol{\epsilon}_t \boldsymbol{\alpha}_1'] = \mathbf{0}.$$

[2] Una variedad lineal $\mathcal{Y}_t$ hace referencia al subespacio cerrado generado por todas las combinaciones lineales finitas de las variables observadas $\{y_1, \ldots, y_t\}$. Intuitivamente, este subespacio contiene toda la información estadística disponible hasta el instante t.

[3] Nótese que, por consistencia con la formulación estándar, se intercambian habitualmente las letras de las matrices de covarianza en función de la especificación concreta del modelo.

4. **Exogeneidad de las Matrices Estructurales:** Las matrices que rigen la dinámica y la medida del sistema $(\mathbf{Z}_t, \mathbf{T}_t, \mathbf{H}_t, \mathbf{Q}_t)$ se consideran deterministas y dadas en la derivación del filtro.[4]

La gran aportación operativa de Kalman (1960), formalizada en su **Teorema 3**, radica en la resolución del problema de almacenamiento y cómputo que lastraba a los enfoques de filtrado clásicos (como el filtro de Wiener). Kalman demostró que la proyección ortogonal del estado sobre toda la historia de observaciones $(\mathcal{Y}_t)$ no requiere recalcularse desde cero; lo que implicaría invertir matrices de datos cada vez más grandes con la llegada de cada nuevo dato temporal.

Bajo este paradigma, el estimador óptimo del estado, denotado como $\mathbf{a}_{t+1|t}$, no se recalcula procesando exhaustivamente la matriz de observaciones históricas en cada instante temporal, sino que se genera de forma recursiva. Kalman demostró que el error de estimación, definido como $\tilde{\boldsymbol{\alpha}}_{t+1|t} = \boldsymbol{\alpha}_{t+1} - \mathbf{a}_{t+1|t}$, evoluciona según su propio sistema dinámico, cuya matriz de covarianza condicional, $\mathbf{P}_{t+1|t} = \mathbb{E}[\tilde{\boldsymbol{\alpha}}_{t+1|t}\tilde{\boldsymbol{\alpha}}'_{t+1|t}]$, concentra toda la incertidumbre del sistema. La minimización de la traza de esta matriz (que representa la pérdida cuadrática esperada) conduce de forma unívoca a las ecuaciones del filtro.

La demostración de Kalman establece que la matriz de ponderación óptima (la Ganancia de Kalman, implícita en su operador Δ^*) y la evolución de la covarianza del error están regidas por una **ecuación matricial no lineal en diferencias**, comúnmente conocida en la teoría de control como la ecuación discreta de Riccati. Adaptando la formulación original de Kalman (Ecuación 32 de su artículo) a la notación que se ha ido siguiendo en el presente trabajo, la recursión para la covarianza del estado toma la forma:

$$\boldsymbol{P_{t+1|t}} = \boldsymbol{T}_t \left\{ \boldsymbol{P}_{t|t-1} - \boldsymbol{P}_{t|t-1}\boldsymbol{Z}'_t \left[\boldsymbol{Z}_t \boldsymbol{P}_{t|t-1} \boldsymbol{Z}'_t + \boldsymbol{H}_t \right]^{-1} \boldsymbol{Z}_t \boldsymbol{P}_{t|t-1} \right\} \boldsymbol{T}'_t + \boldsymbol{R}_t \boldsymbol{Q}_t \boldsymbol{R}'_t. \tag{3.1}$$

Donde $\boldsymbol{P}_{t+1|t}$ y $\boldsymbol{P}_{t|t-1}$ son las matrices de varianzas-covarianzas condicionales del error de estimación del estado, ambas simétricas y de dimensión $m \times m$. La matriz $\boldsymbol{T}_t$, de dimensión $m \times m$, representa la matriz de transición que rige la dinámica autorregresiva de los factores. Por su parte, $\boldsymbol{Z}_t$ es la matriz de medida de dimensión $p \times m$. La matriz $\boldsymbol{H}_t$ $(p \times p)$ contiene la matriz de covarianzas de los errores de medida. Finalmente, la matriz $\boldsymbol{Q}_t$ $(r \times r)$ captura la matriz de covarianzas de las perturbaciones dinámicas que impulsan el estado, cuyo impacto sobre el mismo está mediado por la matriz de selección $\boldsymbol{R}_t$, de dimensión $m \times r$. Es imperativo notar que el término entre corchetes, cuya dimensión es $p \times p$, corresponde a la matriz de covarianza de la innovación $(\boldsymbol{F}_t)$, y su inversión constituye el principal requerimiento computacional del algoritmo.

Esta ecuación es fundamental ya que sustituye la ecuación integral del filtrado de Wiener-Hopf por una recursión matricial resoluble paso a paso. Una vez que $\boldsymbol{P}_{t+1|t}$ es calculada dinámicamente, la especificación explícita del filtro lineal óptimo queda completamente determinada.

Operativamente, y siguiendo la estructura de Kim y Nelson (1999), la resolución de la Ecuación (3.1) y la actualización del estado se desagregan computacionalmente en un ciclo ortogonal de dos fases:

3.2.1.1. Fase I: Predicción

Al inicio del periodo t, condicionado a la variedad lineal de información $\mathcal{Y}_{t-1}$, se proyecta el primer y segundo momento de la distribución del estado a través de la ecuación de transición:

$$\boldsymbol{a}_{t|t-1} = \boldsymbol{T}_t \boldsymbol{a}_{t-1|t-1}. \tag{3.2}$$

$$\boldsymbol{P}_{t|t-1} = \boldsymbol{T}_t \boldsymbol{P}_{t-1|t-1} \boldsymbol{T}'_t + \boldsymbol{R}_t \boldsymbol{Q}_t \boldsymbol{R}'_t.$$

[4] En la práctica, estas matrices son desconocidas y deben ser estimadas en un paso previo, ya sea mediante máxima verosimilitud o el método de componentes principales.

3.2.1.2. Fase II: Actualización mediante el proceso de innovación

Con la llegada de la realización $\mathbf{y}_t$, se calcula la magnitud del vector de **innovaciones**, definido formalmente como el **error de predicción un paso hacia adelante** ($\mathbf{v}_t$). Como señala Kalman (1960), este vector constituye un proceso estocástico ortogonal (ruido blanco) con respecto a la información pasada:

$$\mathbf{v}_t = \mathbf{y}_t - \mathbf{Z}_t \mathbf{a}_{t|t-1},$$

con la matriz de varianza condicional del error de predicción definida como:

$$\mathbf{F}_t = \mathbf{Z}_t \mathbf{P}_{t|t-1} \mathbf{Z}_t' + \mathbf{H}_t.$$

La actualización óptima del estado y la reducción de la incertidumbre se obtienen mediante la proyección de esta innovación ortogonal, ponderada por la Ganancia de Kalman ($\mathbf{K}_t = \mathbf{P}_{t|t-1}\mathbf{Z}_t'\mathbf{F}_t^{-1}$), que define el factor de corrección óptimo:

$$\mathbf{a}_{t|t} = \mathbf{a}_{t|t-1} + \mathbf{K}_t \mathbf{v}_t.$$

$$\mathbf{P}_{t|t} = (\mathbf{I} - \mathbf{K}_t \mathbf{Z}_t)\mathbf{P}_{t|t-1}. \tag{3.3}$$

Corolario 3.1 (Independencia observacional de la covarianza del error)**.** *Dado un sistema dinámico lineal en el Espacio de Estados, la evolución temporal de la matriz de varianzas-covarianzas del error de estimación ($\mathbf{P}_{t|t}$ y $\mathbf{P}_{t+1|t}$), así como la Ganancia de Kalman ($\mathbf{K}_t$), constituyen secuencias deterministas condicionadas estrictamente por las matrices estructurales del sistema ($\mathbf{Z}_t, \mathbf{T}_t, \mathbf{H}_t, \mathbf{Q}_t$) y la condición inicial introducida en la Sección 3.2.1 sobre $\mathbf{P}_1$. Por consiguiente, la reducción analítica de la incertidumbre en el filtro es ortogonal e independiente de las realizaciones del proceso estocástico observable $\{\mathbf{y}_t\}$.*

A diferencia de otros estimadores estadísticos cuya varianza depende de los datos muestrales, en el Filtro de Kalman la matriz $\mathbf{P}_{t|t}$ puede precalcularse en su totalidad antes de observar un solo dato empírico ($\mathbf{y}_t$). Esto garantiza matemáticamente que, si las matrices del sistema están correctamente especificadas y el sistema cumple con las condiciones de observabilidad, el algoritmo iterativo convergerá (incluso en un entorno no estacionario) hacia una reducción óptima del error. A medida que aumenta la longitud del panel temporal T, el sistema refina estructuralmente la extracción de la señal latente, estabilizando la incertidumbre independientemente de los datos observados.

Este conjunto de ecuaciones (3.2)-(3.3) constituye el núcleo que permite estimar los factores latentes del MFD en su representación en el Espacio de Estados.

3.2.2. Estimación de Parámetros y Verosimilitud

Hasta ahora se ha asumido que las matrices estructurales del sistema ($\mathbf{Z}_t, \mathbf{T}_t, \mathbf{H}_t, \mathbf{Q}_t$) son conocidas ex ante. En la práctica, estas estructuras contienen hiperparámetros desconocidos —como las cargas factoriales o los coeficientes autorregresivos del factor— que deben ser estimados.

Una de las grandes ventajas operativas del Filtro de Kalman es que, como subproducto directo de su propia recursión lineal, permite evaluar la función de verosimilitud exacta mediante la descomposición del error de predicción. Tal como detallan Kim y Nelson (1999) y Harvey (1989), la función de log-verosimilitud para una muestra de tamaño T viene dada por la expresión matricial:

$$\ln L(\Theta) = -\frac{1}{2} \sum_{t=1}^{T} \left[p \ln(2\pi) + \ln |\mathbf{F}_t| + \mathbf{v}_t' \mathbf{F}_t^{-1} \mathbf{v}_t \right].$$

Esta formulación permite estimar el vector de parámetros globales Θ del modelo mediante métodos de máxima verosimilitud o algoritmos iterativos como el de *Esperanza-Maximización* (EM), el cual se detallará más en profundidad en la sección correspondiente a los métodos de estimación del MFD.

3.2.2.1. Filtrado vs. Suavizado

Finalmente, es pertinente distinguir, siguiendo a Durbin y Koopman (2012), entre los dos tipos de inferencia que permite el espacio de estados:

- **Filtrado ($\mathbf{a}_{t|t}$):** Representa la proyección ortogonal del vector de estados sobre el espacio lineal generado por las observaciones coetáneas e históricas $\mathcal{Y}_t = \{\mathbf{y}_1, \dots, \mathbf{y}_t\}$, permitiendo actualizar de manera continua la señal latente a medida que se publica nueva información.
- **Suavizado ($\mathbf{a}_{t|T}$):** Utiliza la información de toda la muestra completa $\mathcal{Y}_T = \{\mathbf{y}_1, \dots, \mathbf{y}_T\}$. Mediante una recursión hacia atrás, refina y corrige las estimaciones de los periodos pasados.

Por último, una de las propiedades más valiosas del Filtro de Kalman es su capacidad intrínseca para manejar **paneles desbalanceados o datos ausentes**, una situación recurrente al trabajar con indicadores económicos de coyuntura debido a los calendarios de publicación asíncronos (Angelini, Camba-Méndez, Giannone, Reichlin, y Rünstler, 2008).

Si en el instante t el vector observable $\mathbf{y}_t$ contiene valores perdidos, la dimensión de la ecuación de medida se adapta dinámicamente. Esto se logra premultiplicando $\mathbf{y}_t$ por una matriz de selección $\mathbf{W}_t$ (de dimensión $p_t \times p$), compuesta por ceros y unos, que aísla únicamente las p_t variables disponibles en ese mes concreto. Las matrices del sistema se transforman temporalmente a $\mathbf{y}_t^* = \mathbf{W}_t\mathbf{y}_t$, $\mathbf{Z}_t^* = \mathbf{W}_t\mathbf{Z}_t$ y $\mathbf{H}_t^* = \mathbf{W}_t\mathbf{H}_t\mathbf{W}_t'$, y el filtro opera con total normalidad sobre este subespacio reducido.

En el caso extremo en que todas las observaciones falten en el instante t ($p_t = 0$), el proceso de actualización se omite de forma natural: la innovación y la Ganancia de Kalman se anulan ($\mathbf{K}_t = \mathbf{0}$), y el algoritmo se limita a proyectar su predicción basándose exclusivamente en la dinámica estructural subyacente ($\mathbf{a}_{t|t} = \mathbf{a}_{t|t-1}$ y $\mathbf{P}_{t|t} = \mathbf{P}_{t|t-1}$), generando proyecciones óptimas sin necesidad de interrumpir la ejecución del código.

3.3. El Modelo Factorial Dinámico

Habiendo introducido la representación en el espacio de estados y la mecánica algorítmica del Filtro de Kalman, el siguiente paso es la exposición del Modelo Factorial Dinámico (MFD). Este enfoque generaliza el análisis factorial clásico hacia el dominio de las series temporales, heredando su filosofía fundamental y sus técnicas de interpretación estructural.

Para comprender la naturaleza del MFD, conviene revisar el funcionamiento del análisis factorial y su distinción conceptual respecto al análisis de componentes principales (PCA). El análisis factorial puede interpretarse como un modelo de regresión multivariante que relaciona variables observadas con variables latentes.

Formalmente, dado un conjunto de N variables observadas $\mathbf{X}_t = (X_{1t}, \dots, X_{Nt})'$, se asume que estas se encuentran generadas por r variables latentes inobservables o factores ($\mathbf{f}_t = (f_{1t}, \dots, f_{rt})'$), donde $r \ll N$, mediante la siguiente relación lineal transversal:

$$X_{it} = \lambda_{i1} f_{1t} + \cdots + \lambda_{ir} f_{rt} + u_{it}, \quad i = 1, \dots, N.$$

En esta formulación, los coeficientes λ_{ij}, denominados **cargas factoriales**, miden la sensibilidad o el peso con el que cada factor común incide sobre la variable observada. El término u_{it} representa la perturbación aleatoria, que recoge la dinámica particular y el ruido de medida de X_{it} que no puede ser explicada por las fuerzas comunes, asumiéndose que $\mathbb{E}[u_{it}] = 0$. En consecuencia, la combinación lineal $\sum_{j=1}^{r} \lambda_{ij} f_{jt}$ define la **componente común** de la observación i-ésima de la serie. Se trata, por tanto, de una técnica de reducción de dimensionalidad que busca el número mínimo de dimensiones capaces de explicar la mayor cantidad de información cruzada del panel.

Aunque ambas técnicas comparten el objetivo de comprimir la información, existe una diferencia teórica fundamental. El análisis factorial constituye un modelo estadístico subyacente cuyo objetivo principal es explicar la estructura de las **covarianzas** entre las variables originales. Por el contrario,

el PCA es una transformación ortogonal de carácter algebraico, diseñada para maximizar la **varianza** explicada sin imponer una estructura probabilística previa sobre los términos de error. Sin embargo, bajo ciertas condiciones asintóticas que se detallarán más adelante, el método de componentes principales (PCA) se constituye como el estimador consistente por excelencia de las variables latentes del modelo estadístico factorial.

Más allá de sus propiedades descriptivas, la principal motivación del MFD radica en la resolución del problema de predicción en entornos de datos masivos. Frecuentemente, el objetivo consiste en proyectar una variable de interés a futuro, Y_{t+h}, condicionada a un vector de covariables de alta dimensión $\mathbf{X}_t = (X_{1t}, X_{2t}, \ldots, X_{Nt})'$, un escenario característico del *nowcasting*.

Bajo el enfoque de regresión multivariante tradicional (OLS), proyectar Y_{t+h} sobre la totalidad de $\mathbf{X}_t$ resulta estadísticamente inviable cuando la dimensión transversal N es grande. Por un lado, el error cuadrático medio de la estimación es proporcional a N/T, lo que genera un grave problema de sobreajuste. Por otro lado, si $N > T$, la matriz de covarianzas muestral se vuelve singular, haciendo matemáticamente imposible el cálculo del estimador OLS. Asimismo, las alternativas clásicas de selección de variables basadas en criterios de información como el de Akaike o el Bayesiano exigen la evaluación de 2^N combinaciones, volviéndose computacionalmente inabordables.

El MFD sortea esta «maldición de la dimensionalidad» condensando la información en un número reducido de factores latentes r. Para propósitos de predicción, el modelo dinámico se reescribe en su **representación estática**, apilando los factores y sus retardos en un vector único $\mathbf{F}_t$:

$$\mathbf{X}_t = \mathbf{\Lambda}\mathbf{F}_t + \mathbf{e}_t, \quad t = 1, \ldots, T,$$

donde $\mathbf{F}_t$ es el vector de factores estáticos de dimensión $r \times 1$, $\mathbf{e}_t$ es el vector de perturbaciones aleatorias de media cero, y $\mathbf{\Lambda}$ es la matriz de cargas factoriales de dimensión $N \times r$. Una vez extraída la señal común, la predicción al horizonte h se simplifica a una proyección lineal sobre este espacio de dimensión reducida:

$$Y_{t+h} = \boldsymbol{\alpha}'\mathbf{F}_t + \boldsymbol{\delta}'\mathbf{w}_t + \varepsilon_{t+h}.$$

Los elementos de esta ecuación predictiva se definen de la siguiente forma:

- Y_{t+h}: Variable objetivo adelantada h periodos.
- $\boldsymbol{\alpha}$: Vector de coeficientes $(r \times 1)$ que mide la sensibilidad de la variable dependiente ante los factores comunes.
- $\mathbf{F}_t$: Vector de factores estimados $(r \times 1)$ que condensa la variabilidad compartida del panel.
- $\mathbf{w}_t$: Vector de variables observables exógenas $(m \times 1)$, que habitualmente incluye el intercepto y retardos de la propia variable Y_t mediante un enfoque autorregresivo para capturar persistencias no absorbidas por los factores.
- $\boldsymbol{\delta}$: Vector de parámetros $(m \times 1)$ asociado a las variables observables $\mathbf{w}_t$.
- ε_{t+h}: Error de predicción, asumido ortogonal al conjunto de información disponible en el instante t.

Cabe notar que también puede incluirse Y_t dentro del panel general $\mathbf{X}_t$. Bajo esta especificación, se asume que los factores $\{\mathbf{F}_t\}$ ya absorben toda la información predictiva relevante, aislando el ruido aleatorio y reteniendo exclusivamente los comovimientos dominantes.

Para garantizar la convergencia asintótica del sistema, se asume estrictamente que tanto las variables observables como las latentes son **procesos estocásticos estacionarios, integrados de orden cero,** $I(0)$. La exclusión de raíces unitarias concuerda con la caracterización tradicional del ciclo económico, garantizando que los *shocks* tengan efectos transitorios y no permanentes, tal como subraya Sosa Escudero (1997).

En la práctica, este prerrequisito exige un riguroso procesamiento previo de la información. Las series se transforman mediante diferencias finitas (típicamente diferencias logarítmicas) hasta rechazar la hipótesis de no estacionariedad a través del test de Dickey-Fuller aumentado (véase la Sección A.5.1 del Apéndice A). Posteriormente, los datos se estandarizan para asegurar que posean media cero y varianza unitaria.

3.3.1. Evolución Histórica

Los inicios del MFD en la literatura macroeconómica se remonta a los trabajos seminales de Geweke (1977) y Sargent y Sims (1977), quienes propusieron extensiones de los modelos de factores transversales hacia series temporales, generalizando dinámicamente el clásico modelo de análisis factorial. En su investigación, Sargent y Sims demostraron que un número muy reducido de factores dinámicos era suficiente para explicar una gran fracción de la varianza de las principales variables macroeconómicas estadounidenses, tales como la producción, el empleo y los precios. Posteriormente, Engle y Watson (1981) trasladaron esta metodología al dominio temporal, empleando modelos paramétricos de baja dimensión estimados por Máxima Verosimilitud mediante el Filtro de Kalman.

Sin embargo, esta primera generación de modelos paramétricos sufría de una limitación computacional: la optimización numérica no lineal de la función de verosimilitud se volvía inmanejable conforme aumentaba el número de parámetros. Como consecuencia, su aplicación práctica quedó restringida a escenarios con una baja dimensión (paneles con un número de series N relativamente pequeño).

El salto cualitativo hacia el entorno de paneles de variables grandes se produjo con la segunda generación de estimadores. Stock y Watson (2002a) demostró que, relajando ciertos supuestos paramétricos exactos hacia un "modelo factorial aproximado", era posible aplicar un enfoque no paramétrico basado en el PCA. Este avance teórico permitió tratar con un número N de series mucho mayor. Se demostró que, bajo condiciones donde $N, T \to \infty$, el estimador por Componentes Principales del espacio generado por los factores es consistente y presenta un error cuadrático medio acotado, permitiendo que los factores estimados sean tratados como variables observadas en regresiones subsiguientes. También, en este mismo artículo científico, Stock y Watson (2002a), presentan una alternativa para aplicar PCA a paneles de datos desbalanceados con la presentación del algoritmo EM en el apéndice de su trabajo. Paralelamente, Bai y Ng (2002) y Ahn y Horenstein (2013) desarrollaron criterios para determinar el número óptimo de factores (r) del MFD. Por último, cabe recordar que Bai (2003), de nuevo, derivó la teoría inferencial y las distribuciones límite para estos estimadores lo que fue un gran hallazgo para poder obtener intervalos de confianza de las estimaciones.

La literatura moderna del MFD, liderada por Giannone et al. (2008), representa la tercera generación metodológica, unificando el método de Componentes Principales con la eficiencia del Filtro de Kalman. Este enfoque híbrido emplea el PCA para lidiar con la alta dimensión en una primera etapa, y posteriormente reintroduce los factores en un Espacio de Estados para tratar los datos ausentes y las diferentes frecuencias de publicación mediante el suavizado del Filtro de Kalman.

Cabe destacar que la llegada de la crisis de la COVID-19 ha motivado recientes exploraciones hacia enfoques de estimación bayesiana para incluir una componente estocástica en la volatilidad del modelo. Mediante métodos de simulación como las Cadenas de Markov Monte Carlo (MCMC), los modelos factoriales bayesianos permiten incorporar distribuciones a priori que acomodan parámetros variables en el tiempo y volatilidad estocástica, ofreciendo una implementación práctica más flexible ante *shocks* macroeconómicos, aunque a costa de una carga computacional y complejidad matemática significativamente superior.

3.3.2. Representación del MFD: Dinámica y estática

Una vez presentada una breve cronología historia de la metodología se procede a desarrollar su formulación matemática, la cual se centra en una evolución desde un espacio observable de alta dimensión hacia un subespacio latente de dimensión reducida. En la literatura especializada, esta proyección se parametriza bajo dos estructuras equivalentes: la representación dinámica y la representación estática.

La dicotomía entre ambas representaciones no altera el subespacio común generado, pero condiciona la eficiencia computacional del estimador ante paneles de variables de gran escala (N).

3.3.2.1. Representación Dinámica

La representación dinámica modeliza explícitamente la estructura de retardos que rige la dependencia temporal entre las variables observadas y los factores comunes. Su formulación adopta la estructura clásica de un modelo de espacio de estados lineal:

$$\mathbf{X}_t = \boldsymbol{\lambda}(L)\mathbf{f}_t + \mathbf{e}_t \quad t = 1, \dots, T.$$
$$\mathbf{f}_t = \boldsymbol{\Psi}(L)\mathbf{f}_{t-1} + \boldsymbol{\eta}_t \quad \boldsymbol{\eta}_t \sim \text{i.i.d. } N(\mathbf{0}, \boldsymbol{\Sigma}_\eta),$$

donde $\mathbf{X}_t$ es el vector de observaciones del panel de dimensión $N \times 1$, el vector $\mathbf{f}_t$ de dimensión $q \times 1$ contiene los factores dinámicos, y $\boldsymbol{\lambda}(L)$ representa la matriz de polinomios de retardo de dimensión $N \times q$ que define las cargas factoriales dinámicas. El término $\mathbf{e}_t = (e_{1t}, \dots, e_{Nt})'$ constituye el vector de perturbaciones aleatorias de dimensión $N \times 1$. El producto $\boldsymbol{\lambda}(L)\mathbf{f}_t$ determina la componente común del sistema, mientras que la transición del estado latente sigue un proceso autorregresivo vectorial gobernado por el polinomio matricial $\boldsymbol{\Psi}(L)$. Asimismo, se impone la condición de ortogonalidad entre las innovaciones de los factores y las perturbaciones aleatorias para todos los retardos y adelantos, de modo que $\mathbb{E}[\mathbf{e}_t\boldsymbol{\eta}'_{t-k}] = \mathbf{0}, \forall k$.

Bajo los supuestos de normalidad multivariante y ortogonalidad del vector de errores $\mathbf{e}_t$, el predictor lineal óptimo a un paso de la serie i-ésima, calculado mediante el Filtro de Kalman, se reduce a la proyección sobre el espacio de los factores y la propia inercia histórica de la variable:

$$\mathbb{E}[X_{it+1}|\mathbf{X}_t, \mathbf{f}_t, \mathbf{X}_{t-1}, \mathbf{f}_{t-1}, \dots] = \boldsymbol{\alpha}_i(L)\mathbf{f}_t + \delta_i(L)X_{it}.$$

Esto demuestra de manera analítica que el subespacio factorial absorbe asintóticamente la totalidad de la información cruzada compartida por el panel de indicadores.

3.3.2.2. Representación Estática y Transformación de Markov

La estimación paramétrica directa de los polinomios de retardo $\boldsymbol{\lambda}(L)$ mediante máxima verosimilitud tiende a divergir numéricamente cuando la dimensión transversal N es elevada. Para sortear esta restricción algorítmica, el modelo se reescribe apilando los factores dinámicos contemporáneos y sus correspondientes retardos hasta un orden finito p. De este modo se construye un vector de estado ampliado $\mathbf{F}_t = (\mathbf{f}'_t, \mathbf{f}'_{t-1}, \dots, \mathbf{f}'_{t-p})'$ de dimensión $r \times 1$, donde $q(p+1) = r \geq q$. Esta transformación matemática anula la estructura dinámica explícita en la ecuación de medida, dando lugar a la **representación estática**.

Considérese el vector $\mathbf{X}_t$ de dimensión $N \times 1$ que unifica los datos observados de la sección transversal para un instante temporal t, con $t = 1, \dots, T$. La estructura del modelo se redefine mediante el siguiente sistema dinámico:

$$\mathbf{X}_t = \boldsymbol{\Lambda}\mathbf{F}_t + \mathbf{e}_t.$$
$$\mathbf{F}_t = \boldsymbol{\Phi}\mathbf{F}_{t-1} + \mathbf{G}\boldsymbol{\eta}_t.$$

La equivalencia algebraica entre ambas representaciones queda expuesta detalladamente en Stock y Watson (2016). Suponiendo un escenario simplificado con $q = 1$ factor dinámico, un único retardo en la carga factorial y un proceso VAR(2) en la ecuación de transición ($f_t = \Psi_1 f_{t-1} + \Psi_2 f_{t-2} + \eta_t$), la parametrización estática exige definir $r = 2$ factores, lo que resulta en el sistema markoviano de bloques:

$$\mathbf{X}_t = \begin{pmatrix} \boldsymbol{\lambda}_0 & \boldsymbol{\lambda}_1 \end{pmatrix} \begin{pmatrix} f_t \\ \\ f_{t-1} \end{pmatrix} + \mathbf{e}_t = \boldsymbol{\Lambda}\mathbf{F}_t + \mathbf{e}_t.$$

$$\begin{pmatrix} f_t \\ f_{t-1} \end{pmatrix} = \begin{pmatrix} \Psi_1 & \Psi_2 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} f_{t-1} \\ f_{t-2} \end{pmatrix} + \begin{pmatrix} 1 \\ 0 \end{pmatrix} \eta_t \implies \mathbf{F}_t = \mathbf{\Phi}\mathbf{F}_{t-1} + \mathbf{G}\boldsymbol{\eta}_t.$$

La ventaja fundamental de esta representación radica en que permite estimar el subespacio abarcado por el vector de factores $\mathbf{F}_t$ a través del análisis de componentes principales (PCA) basándose de manera exclusiva en la varianza transversal, eludiendo la necesidad de imponer una modelización paramétrica previa sobre la matriz de transición $\mathbf{\Phi}$.

3.3.2.3. Relajación de Supuestos: Del MFD Exacto al Aproximado

La viabilidad de aplicar un estimador basado en Componentes Principales (PCA) o en técnicas de Máxima Verosimilitud sobre la ecuación de medida estática de un Modelo Factorial Dinámico está condicionada por el comportamiento y la estructura de la sección transversal de sus perturbaciones. Esta relación se formaliza mediante la matriz de varianzas-covarianzas de los errores, definida como $\mathbf{\Sigma}_e = \mathbb{E}[\mathbf{e}_t\mathbf{e}_t']$, una matriz cuadrada de dimensión $N \times N$ cuyos elementos diagonales σ_{ii} representan la varianza particular de cada indicador y cuyos elementos fuera de la diagonal σ_{ij} capturan las covarianzas contemporáneas cruzadas entre los residuos de las distintas variables. En la literatura macroeconómica, la naturaleza de esta matriz determina la división entre dos paradigmas estructurales:

- **MFD Exacto:** Es el enfoque clásico y más restrictivo. Asume que la matriz $\mathbf{\Sigma}_e$ es estrictamente diagonal, lo que impone una ausencia total de correlación transversal contemporánea entre los errores de los distintos indicadores del panel:

$$\mathbf{\Sigma}_e = \begin{pmatrix} \sigma_1^2 & 0 & \dots & 0 \\ 0 & \sigma_2^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma_N^2 \end{pmatrix} \implies \mathbb{E}[e_{it}e_{jt}] = 0, \quad \forall i \neq j.$$

  Bajo esta premisa, se asume que el vector de factores comunes $\mathbf{F}_t$ absorbe de forma absoluta e íntegra la totalidad del comovimiento existente en el panel de datos. Esto implica que, una vez condicionado el sistema al conocimiento de las variables latentes, el resto de las series explicativas se vuelven mutuamente independientes, careciendo de cualquier poder de retroalimentación o capacidad predictiva cruzada entre sí. Aunque esta restricción simplifica la optimización computacional en muestras pequeñas, resulta excesivamente rígida e irreal en la econometría aplicada de coyuntura. Al trabajar con paneles reales de indicadores, la presencia de eslabonamientos productivos sectoriales, encadenamientos de suministro o regulaciones institucionales compartidas tiende a generar dependencias mutuas residuales entre las perturbaciones específicas de las variables.

- **MFD Aproximado:** Formalizado originalmente por Chamberlain y Rothschild (1983) para entornos estáticos y extendido formalmente a las series temporales de alta dimensión por Stock y Watson (2002b), este enfoque flexibiliza el supuesto clásico al admitir que la matriz de covarianzas $\mathbf{\Sigma}_e$ no sea diagonal, tolerando de este modo una estructura de dependencia transversal y serial "débil" entre las perturbaciones aleatorias.

  Para salvaguardar la identificabilidad matemática del factor común y garantizar que el algoritmo no confunda el ruido sectorial local con la señal del ciclo económico general, el concepto de **correlación transversal débil** se define mediante el comportamiento de los autovalores (eigenvalues) de las matrices del sistema cuando la dimensión transversal del panel tiende a infinito ($N \to \infty$):

1. *Acotación geométrica del ruido aleatorio:* Sea $\lambda_{\max}(\mathbf{\Sigma}_e)$ el mayor autovalor de la matriz de varianzas-covarianzas de los errores. Se exige estrictamente que permanezca acotado por una constante real fija M ante el incremento transdimensional de la sección cruzada:

$$\lambda_{\max}(\mathbf{\Sigma}_e) \leq M < \infty \quad \forall N.$$

Una condición matemática equivalente y de gran utilidad econométrica para interpretar esta restricción consiste en exigir la sumabilidad absoluta de las covarianzas cruzadas fuera de la diagonal para cualquier variable i:

$$\max_{i \leq N} \sum_{j=1}^{N} |\mathbb{E}[e_{it} e_{jt}]| \leq M < \infty.$$

Esta propiedad garantiza que las relaciones de dependencia entre los residuos afecten únicamente a subgrupos limitados de variables (clústeres sectoriales) y tiendan a disiparse promedialmente a medida que el tamaño del panel se expande de forma masiva.

2. *Divergencia asintótica de la componente común:* En contraposición a la naturaleza local del ruido, las fuerzas macroeconómicas comunes deben impactar globalmente en el sistema. Por consiguiente, se requiere que el r-ésimo autovalor de la matriz de covarianzas de la componente común, generada a partir de la matriz de cargas factoriales como $\mathbf{\Lambda\Lambda'}$, diverja asintóticamente a una tasa de orden N:

$$\lambda_r(\mathbf{\Lambda'\Lambda}) \to \infty \quad \text{con} \quad \lim_{N\to\infty} \frac{\mathbf{\Lambda'\Lambda}}{N} = \mathbf{D} > 0,$$

donde $\mathbf{D}$ constituye una matriz límite de parámetros de dimensión $r \times r$, estrictamente definida positiva.

La interacción de ambas condiciones asintóticas es la que legitima el uso de los modelos factoriales dinámicos aproximados en el análisis de coyuntura real. Al expandir el tamaño de la sección transversal, la varianza explicada por la señal macroeconómica del ciclo diverge a una tasa proporcional a N, mientras que la varianza acumulada por las perturbaciones aleatorias se mantiene estancada bajo la frontera de la cota superior M.

Como consecuencia de esta asimetría geométrica, las correlaciones débiles remanentes se cancelan mutuamente por un efecto de promedio amparado en la Ley de los Grandes Números para datos dependientes. El ruido local se diluye en el agregado, permitiendo al investigador incorporar de forma segura agrupaciones de variables muy conectadas en la realidad productiva (v.g., los acoplamientos entre el Índice de Producción Industrial y las ventas de las grandes empresas, o las dinámicas logísticas cruzadas en las aduanas) sin temor a sesgar la trayectoria del factor latente ni alterar la consistencia de las proyecciones del *nowcasting*.

3.3.2.4. El Problema de la Identificabilidad: Rotación e Indeterminación

La estimación de un MFD se enfrenta a un desafío estructural intrínseco derivado de la inobservabilidad de los factores latentes. Si bien el espacio lineal abarcado por la trayectoria temporal de las variables latentes $\{\mathbf{F}_t\}_{t=1}^{T}$ es identificable, el modelo carece de una solución única para sus parámetros a menos que se impongan restricciones matemáticas de identificación. Este fenómeno, denominado *el problema de identificabilidad*, implica la existencia de infinitas combinaciones alternativas de matrices de cargas ($\mathbf{\Lambda}$) y vectores de factores ($\mathbf{F}_t$) que devuelven exactamente la misma función de verosimilitud para el panel de datos observados $\mathbf{X}_t$.

Para formalizar esta indeterminación algebraica, considérese la representación estática del modelo $\mathbf{X}_t = \mathbf{\Lambda F}_t + \mathbf{e}_t$. Si se define una matriz de transformación lineal $\mathbf{H}$ de dimensión $r \times r$ que sea invertible y no singular, el sistema puede reescribirse de forma equivalente como:

$$\mathbf{X}_t = (\mathbf{\Lambda H}^{-1})(\mathbf{H F}_t) + \mathbf{e}_t = \tilde{\mathbf{\Lambda}}\tilde{\mathbf{F}}_t + \mathbf{e}_t,$$

donde las nuevas cargas factoriales $\tilde{\mathbf{\Lambda}} = \mathbf{\Lambda}\mathbf{H}^{-1}$ y los factores transformados $\tilde{\mathbf{F}}_t = \mathbf{H}\mathbf{F}_t$ reproducen con exactitud la dinámica observable de la serie. La naturaleza geométrica de esta indeterminación varía sustancialmente según el número de factores determinados en el sistema.

- **Caso de un único factor ($r = 1$): Escala y signo**

  Cuando la estructura del modelo se restringe a la extracción de un único factor común (asimilable conceptualmente al pulso general de la actividad económica), la matriz de rotación $\mathbf{H}$ colapsa a un escalar multiplicativo h. En este escenario, la rotación geométrica es inexistente debido a que un espacio unidimensional no puede rotar sobre sí mismo; sin embargo, persisten dos tipos de indeterminación matemática:

  1. **Indeterminación de escala:** El algoritmo es incapaz de discriminar entre un factor latente de alta volatilidad ponderado por cargas factoriales reducidas, y un factor de baja varianza asociado a cargas de gran magnitud ($h \in \mathbb{R}$). Como se ilustra esquemáticamente en la Figura 3.2, un factor original $\mathbf{F}_1$ puede ser representado mediante un vector transformado $\mathbf{F}_1'$ que duplique su longitud escalar. Para que la igualdad del sistema se preserve, esta dilatación en la magnitud del factor requiere una contracción proporcional de las cargas factoriales de la matriz $\mathbf{\Lambda}$, dividiendo su valor por dos.

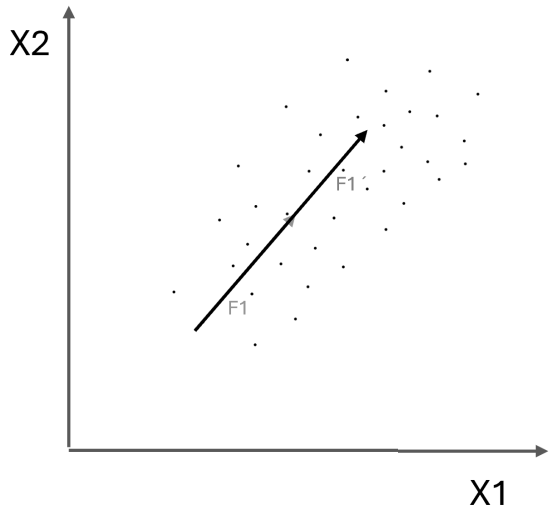


Figura 3.2: Indeterminación de escala en el modelo de factor único ($r = 1$).Elaboración propia.

.

  2. **Indeterminación de signo:** El factor común puede interpretarse indistintamente en sentido de expansión o de recesión mediante la simple multiplicación por un escalar $h = -1$, sin alterar en absoluto la bondad del ajuste del modelo. Como se ilustra en la Figura 3.3, esta dualidad implica que la trayectoria geométrica de la señal latente mantiene la misma distancia euclídea respecto a las variables observables del panel, variando únicamente la orientación vectorial de su signo.

  Una opción para resolver esta problemática consiste en la imposición de restricciones normativas sobre los parámetros. Una alternativa habitual radica en normalizar la varianza del factor a la unidad ($\text{Var}(\mathbf{F}_t) = 1$), lo que determina de forma unívoca la longitud del vector en su subespacio. Para resolver la dirección del signo, se impone simultáneamente una restricción de positividad sobre la carga factorial de una variable macroeconómica de referencia macroeconómica clara

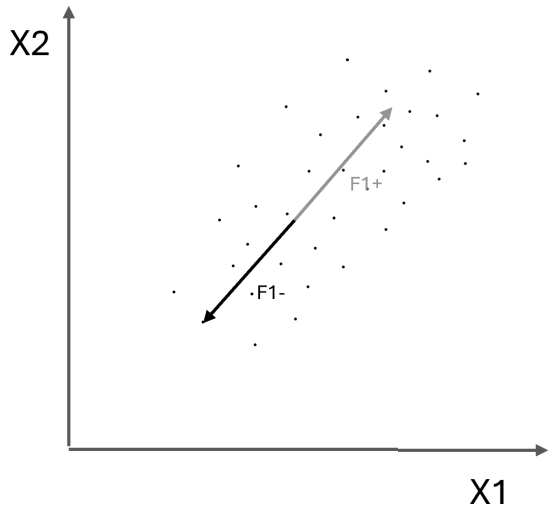


Figura 3.3: Indeterminación de signo en el modelo de factor único ($r = 1$). Elaboración propia.

.

(p.ej., el PIB o el IPI), forzando al factor a evolucionar en consonancia con el ciclo económico real. Otra estrategia formal para subsanar la indeterminación de escala consiste en la fijación de una carga unitaria. Mediante este enfoque, se selecciona una variable de alta relevancia dentro del panel y se establece su coeficiente de carga estrictamente igual a uno, obligando matemáticamente al factor latente a adoptar la misma escala de medida y magnitud que dicha variable de interés.

- **Caso de múltiples factores ($r > 1$): Rotación e hiperplanos**

  Cuando el modelo incorpora más de un factor común, surge la indeterminación por rotación factorial. En este escenario, las variables latentes definen un subespacio geométrico (un plano bidimensional si $r = 2$) inmerso en la nube de datos de dimensión N. La matriz de transformación $\mathbf{H}$, al ser invertible, permite que los ejes factoriales giren libremente dentro de dicho subespacio sin alterar la distancia euclídea de las observaciones respecto al plano común. Esta problemática se ilustra en la Figura 3.4 siendo F1 y F2 el factor 1 y 2, respectivamente.

  Geométricamente, para el caso de dos factores, la matriz $\mathbf{H}$ adopta la forma de una matriz de rotación ortogonal parametrizada por un ángulo θ:

$$\mathbf{H} = \begin{pmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{pmatrix}.$$

  Dado que existe un continuo infinito de ángulos posibles, coexisten infinitas maneras de proyectar los ejes de los factores dentro del subespacio, devolviendo representaciones numéricas distintas para una misma realidad económica.

  Sin la imposición de restricciones normativas, el algoritmo de estimación es incapaz de fijar la orientación de los ejes, lo que impide la convergencia numérica y obstaculiza la interpretación económica de los factores. Para solucionar esta indeterminación, la literatura econométrica recurre a una serie de normalizaciones que anulan los grados de libertad de giro de la matriz $\mathbf{H}$.

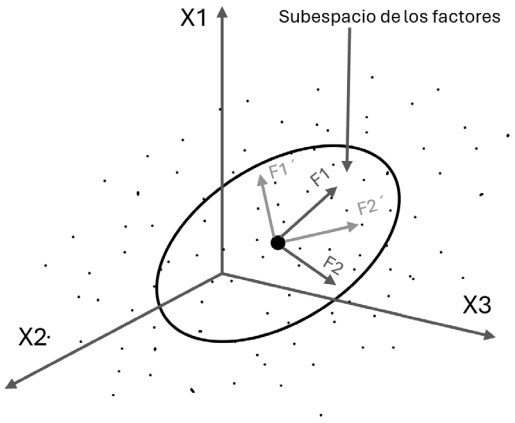


Figura 3.4: Indeterminación rotacional en el modelo de dos factores ($r = 2$).Elaboración propia.

.

1. **Normalización por componentes principales** Para resolver la indeterminación rotacional definida por $\tilde{\mathbf{\Lambda}} = \mathbf{\Lambda}\mathbf{H}^{-1}$ y $\tilde{\mathbf{F}}_t = \mathbf{H}\mathbf{F}_t$, se imponen restricciones asintóticas que fuerzan una solución única basada en la estructura de autovalores del panel de datos:
   - **Regla del eje mayor y ordenación de varianza:** Se exige que el primer factor se oriente hacia la dirección de máxima variabilidad del subespacio. Matemáticamente, esto requiere el cumplimiento de dos condiciones simultáneas: en primer lugar, la matriz $\mathbf{\Lambda}'\mathbf{\Lambda}$ debe ser estrictamente diagonal; en segundo lugar, sus elementos deben estar ordenados de manera decreciente ($\gamma_1 > \gamma_2 > \ldots > \gamma_r$). Esta doble restricción matemática garantiza que los factores latentes queden unívocamente ordenados según su capacidad para explicar la varianza común del panel.
   - **Ortogonalidad estructural:** Se impone que los factores comunes sean ortogonales entre sí en el tiempo, de modo que $\mathbb{E}[\mathbf{F}_t\mathbf{F}_t'] = \mathbf{\Sigma}_F$, donde $\mathbf{\Sigma}_F$ es una matriz diagonal. Geométricamente, esta condición fuerza a que los ejes factoriales mantengan una separación perpendicular de 90 grados, asegurando que capturen dimensiones no solapadas del ciclo económico.
   - **Restricción de escala asintótica:** Se fija la escala de la matriz de cargas mediante la condición identificadora $\frac{\mathbf{\Lambda}'\mathbf{\Lambda}}{N} = \mathbf{I}_r$. Esta restricción es crucial en entornos de alta dimensión ($N \to \infty$), ya que evita que la norma de las cargas factoriales crezca indefinidamente al incorporar nuevas series, asegurando la consistencia asintótica del estimador PCA.
2. **Normalización mediante restricciones estructurales o de exclusión**
   Como detallan Stock y Watson (2016), una alternativa para fijar la rotación consiste en vincular *a priori* un factor con la dinámica de un grupo específico de variables observables emparentadas (p.ej., indicadores del sector consumo). Esto se instrumenta imponiendo restricciones de exclusión, fijando estrictamente a cero las cargas factoriales de aquellas variables ajenas al sector de referencia, lo que estabiliza la orientación de los ejes mediante criterios de teoría económica.

3.3.2.5. Consideraciones estructurales del modelo

Las referencias bibliográficas con las que se ha apoyado la redacción de esta sección también sugieren tres cuestiones fundamentales que han descubierto en sus estudios empíricos y que pueden interesar conocer de cara a la aplicación empírica del modelo:

1. Dimensión del panel (N) y ruido: Aunque la teoría asintótica de Bai (2003) requiere $N, T \to \infty$, la adición indiscriminada de variables incrementa la varianza de la perturbación aleatoria y puede violar las condiciones del MFD aproximado (Bai y Ng, 2008). Es por ello que debe llevarse a cabo criterios de selección de variables y no añadir de forma indiscriminada variables al modelo. Sumado a esto, Stock y Watson (2002b) mostraron que la mejora del ajuste cuando N supera las 50 variables es muy reducido y no compensa el incremento de complejidad al modelo.

2. Estacionariedad frente a dinámicas integradas: Mientras que aproximaciones metodológicas como la de Harvey (1989) permiten modelizar paneles donde las series presentan raíces unitarias $I(1)$ mediante la inclusión de tendencias estocásticas comunes, la implementación del PCA estándar sobre la que pivota el análisis de grandes paneles moderno (Giannone et al., 2008; Stock y Watson, 2002b) exige **estacionariedad débil,** $I(0)$. Para evitar que la presencia de raíces unitarias distorsione la estructura de covarianzas transversales, las series macroeconómicas de nivel se someten a una transformación en búsqueda de cumplir esta condición.

3. Robustez ante Inestabilidad Estructural: La asunción paramétrica de constancia en $\mathbf{\Lambda}$ es inherentemente frágil en macroeconomía debido a cambios de régimen. Sin embargo, Stock y Watson (2009) demuestran analíticamente que la estimación del espacio latente mediante PCA es asintóticamente robusta frente a *breaks* estructurales, siempre que la magnitud de la ruptura esté acotada a $O(N^{-1/2})$ o que los cambios en las cargas sigan procesos de paseo aleatorio independientes. Esta propiedad justifica la estimación de los factores sobre toda la muestra disponible, blindando el *nowcast* frente a inestabilidades moderadas de las series individuales.

3.3.3. Métodos de estimación del modelo

Una vez formalizadas las distintas representaciones del Modelo Factorial Dinámico (MFD), se prosigue con la estimación del subespacio inobservable generado por los factores latentes. Dado que los factores F_t no son directamente observables, el problema de estimación se desplaza hacia la identificación del espacio definido por estos y su interacción con el panel de variables X_t. Históricamente, la literatura ha transitado desde sistemas de baja dimensión (N pequeño) a partir de una representación del MFD como un modelo del Espacio de Estados lineal ajustado de forma paramétrica en el dominio temporal a través de la estimación por Máxima Verosimilitud (ML) y el Filtro de Kalman, hacia otros entornos de alta dimensión ($N \gg T$). En este nuevo paradigma, el número de series N puede ser superior al número de observaciones temporales T, lo que invalida los métodos de optimización no lineal tradicionales y motiva el uso de técnicas no paramétricas basadas en PCA para estimar el espacio generado de los factores.

Finalmente, se examinará la vanguardia de los MFD hacia el enfoque bayesiano, que ya se comentó durante esta sección, una transición precipitada por coyunturas *sui generis* en la dinámica económica reciente- en concreto la crisis sufrida por la pandemia del COVID-19- que han exigido una adaptación técnica producto del aumento de las volatilidades de la series y a la pérdida de relaciones históricas entre variables.

3.3.3.1. Enfoque paramétrico: Máxima Verosimilitud en el Espacio de Estados

El enfoque paramétrico considera al MFD como un sistema estocástico completamente gobernado por un vector de parámetros $\boldsymbol{\theta} = \{\mathbf{\Lambda}, \mathbf{\Phi}, \mathbf{\Sigma}_\xi, \mathbf{\Sigma}_\eta\}$. Bajo esta perspectiva, el modelo se formula en su representación de **Espacio de Estados (SS)**, cuya estructura permite capturar la dinámica de los factores y su impacto en las variables observadas de forma simultánea. Este método constituye el estándar de eficiencia estadística cuando la dimensión transversal N es pequeña y fija, permitiendo una estimación conjunta mediante métodos de máxima verosimilitud.

En este enfoque, el Filtro de Kalman desempeña un papel dual. Por un lado, actúa como el algoritmo recursivo que permite evaluar la función de verosimilitud; por otro, funciona como el estimador óptimo de mínima varianza posible entre los estimadores insesgados de los factores, bajo las condiciones iniciales descritas en la Sección 3.2.1 de este mismo capítulo. Siguiendo el proceso basado en la

descomposición del error de predicción (Harvey, 1989), la función de log-verosimilitud se define como:

$$\ln L(\boldsymbol{\theta}; \mathbf{Y}) = -\frac{NT}{2}\ln(2\pi) - \frac{1}{2}\sum_{t=1}^{T}\left(\ln|\mathbf{F}_t| + \boldsymbol{\nu}_t'\mathbf{F}_t^{-1}\boldsymbol{\nu}_t\right). \tag{3.4}$$

Donde $\boldsymbol{\nu}_t$ representa el vector de innovaciones y $\mathbf{F}_t$ su correspondiente matriz de varianza-covarianza.

El Filtro de Kalman recorre la muestra cronológicamente, actualizará en cada paso t la estimación sobre el factor $\mathbf{F}_t$ tras observar el nuevo dato $\mathbf{y}_t$. Dada la no linealidad de $\ln L(\boldsymbol{\theta})$ respecto a los parámetros, la obtención del estimador $\hat{\boldsymbol{\theta}}_{ML}$ requiere de algoritmos de optimización numérica que logren un equilibrio entre estabilidad y eficiencia computacional:

1. **El Algoritmo BFGS**: Para evitar el coste computacional $O(k^3)$, donde k representa el número de parámetros a estimar, asociado al cálculo y la inversión de la matriz Hessiana en cada iteración del método de Newton-Raphson, se recurre frecuentemente a algoritmos *Quasi-Newton*, siendo el BFGS (*Broyden-Fletcher-Goldfarb-Shanno*) el más extendido. El BFGS construye una aproximación iterativa de la inversa de la Hessiana ($\mathbf{B}^{-1}$) basándose en la ecuación de la secante y los cambios observados en el gradiente ($\nabla\ell$). Utiliza actualizaciones de rango 2 para asegurar que la matriz de curvatura sea simétrica y definida positiva, presentando una tasa de convergencia superlineal que evita la inestabilidad de los métodos de segundo orden en zonas de baja curvatura o mesetas.

2. **Template Model Builder (TMB)**: una extensión o modificación de este método es el framework TMB (Kristensen, Nielsen, Berg, Skaug, y Bell, 2016), que integra tres pilares tecnológicos:

   - *Diferenciación Automática (AD):* Mediante librerías del lenguaje de programación C++, calcula gradientes y Hessianas de forma más eficiente.
   - *Aproximación de Laplace:* Útil para integrar factores fuera de la verosimilitud en problemas de optimización tratables.
   - *Matrices Dispersas:* Explota la estructura de ceros en las matrices de transición, permitiendo escalar el modelo a dimensiones inalcanzables para *solvers* estándar.

3. **Algoritmo Expectación-Maximización (EM)**: Introducido por Watson y Engle (1983), ofrece una alternativa robusta al tratar los factores latentes $\mathbf{F}_t$ como datos ausentes, simplificando la maximización mediante un proceso iterativo de dos pasos:

   - *Paso E (Expectation)*: Utiliza el Suavizador de Kalman para calcular la esperanza condicional de la log-verosimilitud completa.
   - *Paso M (Maximization)*: Actualiza el vector $\boldsymbol{\theta}^{(i+1)}$ mediante regresiones lineales simples de Mínimos Cuadrados Ordinarios (MCO).

El enfoque paramétrico mediante Máxima Verosimilitud (MLE) se enfrenta a restricciones teóricas y operativas severas cuando la dimensión (N) se expande. Estas limitaciones se pueden sistematizar en tres ejes fundamentales:

1. **Crecimiento Paramétrico y Pérdida de Identificación**: el número total de parámetros a estimar depende de la arquitectura del modelo y del supuesto sobre los errores aleatorios. El crecimiento de k sigue dos trayectorias posibles según la flexibilidad que otorguemos al sistema:

   (MFD Exacto): En la especificación estándar, se asume que los errores aleatorios son independientes entre sí (matriz $\boldsymbol{\Sigma}_e$ diagonal). En este caso, k se compone de $N \times r$ cargas factoriales, N varianzas de los errores y, opcionalmente, N coeficientes autorregresivos para los errores. Por tanto, $k \approx N(r+2)$, lo que implica un crecimiento lineal respecto al número de series, sea $O(N)$.

   (MFD Aproximado sin restricciones): Si se pretende modelizar la estructura de dependencia transversal de los errores mediante una matriz de covarianza completa (no diagonal), el número

de parámetros de dicha matriz escala como $\frac{N(N+1)}{2}$. En este punto, k crece de forma cuadrática, $O(N^2)$, saturando la capacidad del modelo para identificar los parámetros a partir de una muestra temporal T finita (problema de $N \gg T$).

2. **Complejidad Computacional y Saturación de la Inversión**: el cuello de botella de los métodos de optimización tradicionales (como BFGS o Newton-Raphson) reside en la actualización de los parámetros en cada iteración. Estos algoritmos requieren calcular o aproximar la matriz Hessiana (la matriz de segundas derivadas de la log-verosimilitud). La inversión de esta matriz de tamaño $k \times k$, necesaria para determinar la dirección de búsqueda del máximo, tiene un coste computacional de orden cúbico, $(O(k^3))$. Mientras que en el BFGS el coste es prohibitivo para $N > 30$ (es el límite aproximado de series que se indica en la literatura especializada) debido al ya mencionado $O(k^3)$, el algoritmo EM consigue una ventaja disruptiva al operar con una complejidad de $O(N)$ por iteración en su Paso M, ya que estima las cargas variable por variable. No obstante, la naturaleza iterativa del EM implica que, aunque cada paso sea rápido, el número total de pasos para converger puede ser muy elevado en dimensiones altas.

3. **Geometría de la Verosimilitud e Inestabilidad Numérica**: a medida que el espacio de parámetros crece, la función de log-verosimilitud se vuelve extremadamente compleja de buscar su máximo:

   - Mesetas y Singularidad: En alta dimensión, la superficie de verosimilitud tiende a volverse plana en múltiples direcciones (gradiente casi nulo). Esto provoca que la matriz Hessiana sea numéricamente singular o esté mal condicionada, lo que impide que los algoritmos identifiquen un máximo global único y genera estimadores con varianzas asintóticas enormes.
   - Multimodalidad: Con un N elevado, aumentan las probabilidades de que existan múltiples máximos locales. Según Stock y Watson (2016), esto deriva en que los algoritmos converjan a soluciones subóptimas dependiendo críticamente de los valores iniciales, una debilidad que incluso el algoritmo EM comparte si no se inicializa mediante métodos consistentes como el PCA.

3.3.3.2. Enfoque no paramétrico: Análisis de Componentes Principales (PCA)

En escenarios caracterizados por una elevada dimensión transversal del panel donde $N \gg T$, los métodos de estimación basados en la maximización iterativa de la función de verosimilitud suelen incurrir en la saturación de los recursos computacionales o en la inestabilidad de la superficie de optimización. Ante esta restricción, la literatura -encabezada por Stock y Watson (2002b) y formalizada analíticamente por Bai (2003)- propone el uso del PCA como un método determinista y no paramétrico para la extracción de los factores latentes. A diferencia de los algoritmos basados en el gradiente, el PCA elude la necesidad de modelizar explícitamente las leyes de transición del sistema en una primera etapa, obteniendo los factores mediante una solución algebraica de forma cerrada.

Partiendo de la representación estática del MFD, $\mathbf{X}_t = \mathbf{\Lambda F}_t + \mathbf{e}_t$, el estimador de componentes principales busca minimizar la suma de los residuos al cuadrado promediada a través de la resolución del siguiente problema de optimización de mínimos cuadrados ordinarios:

$$V(k, \hat{\mathbf{F}}^k) = \min_{\mathbf{\Lambda}, \mathbf{F}^k} \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} (x_{it} - \boldsymbol{\lambda}_i' \mathbf{F}_t^k)^2.$$

Sujeto a las restricciones de identificación necesarias para fijar la escala y la rotación del subespacio factorial, introducidas en la Sección 3.3.2.4. Matemáticamente, las columnas del vector de factores estimados $\hat{\mathbf{F}}$ se obtienen a partir de los autovectores correspondientes a los r mayores autovalores de la matriz de varianza-covarianza muestral de los datos, definida como $\hat{\mathbf{\Sigma}}_X = \frac{1}{T} \sum_{t=1}^{T} \mathbf{X}_t \mathbf{X}_t'$. Esta propiedad convierte un problema de inferencia de variables latentes en un ejercicio de descomposición

espectral, garantizando una solución única y globalmente óptima bajo los supuestos de identificación impuestos.

La transición hacia este enfoque permite relajar los supuestos restrictivos del MFD exacto -donde se exige la independencia de los errores aleatorios- para operar bajo el marco del **Modelo Factorial Aproximado**, cuyas características fueron previamente introducidas en la Sección 3.3.2.3. Siguiendo a Chamberlain y Rothschild (1983), el PCA resulta consistente incluso en presencia de una correlación transversal y serial "débil" en $\mathbf{e}_t$. La validez estadística de este estimador reside en sus propiedades de convergencia asintótica: Bai (2003) demuestra que cuando $N, T \to \infty$, los factores estimados $\hat{\mathbf{F}}_t$ convergen al espacio generado por los factores verdaderos ($\mathbf{HF}_t$) a una tasa dada por:

$$\text{mín}\{\sqrt{N}, T\} \cdot (\hat{\mathbf{F}}_t - \mathbf{HF}_t) = O_p(1).$$

Esta propiedad garantiza que los factores puedan ser tratados como variables observadas en etapas posteriores del análisis, como la estimación de las cargas factoriales mediante regresiones por Mínimos Cuadrados Ordinarios (MCO), sin incurrir en sesgos por error de medición en el límite asintótico. Este estudio de la convergencia y las propiedades de las estimaciones se detallará de forma más exhaustiva en la Sección 3.3.5.

Desde una perspectiva de eficiencia computacional, el uso del PCA presenta una ventaja frente al coste $O(k^3)$ -que asintóticamente equivale a $O(N^3)$- de los métodos de gradiente basados en la matriz Hessiana, o frente a la naturaleza iterativa del algoritmo EM, cuyo coste por iteración escala de forma lineal $O(N)$. La complejidad del cálculo de la descomposición en autovalores es de orden $O(\text{mín}\{N^2T, NT^2\})$, lo que permite el tratamiento de paneles masivos frente a los métodos iterativos previamente introducidos. No obstante, el PCA presenta limitaciones críticas para la práctica del *nowcasting*: en primer lugar, su naturaleza atemporal le impide explotar la dinámica autorregresiva de los factores durante su estimación; en segundo lugar, su exigencia de paneles balanceados obliga a truncar la muestra ante la presencia de bordes dentados, desechando la información más reciente de los indicadores de alta frecuencia, perdiendo la estimación temprana de la actividad económica, que es el objetivo último que se persigue en el presente proyecto y una de las principales ventajas de utilizar el MFD.

3.3.3.3. Métodos híbridos de estimación

Una vez evaluada la dicotomía existente entre la eficiencia estadística del enfoque paramétrico convencional y la robustez computacional del análisis no paramétrico por componentes principales, la literatura macroeconómica moderna ha convergido hacia el desarrollo de técnicas de naturaleza híbrida.

Cabe señalar que esta última vertiente constituye el núcleo metodológico sobre el cual se articula la totalidad de la aplicación empírica del presente proyecto. La elección de esta arquitectura responde a su capacidad para procesar paneles de alta dimensión conservando la flexibilidad del espacio de estados, requisito indispensable para el manejo de las perturbaciones autorregresivas y el flujo asíncrono de datos de la economía española. Al ser la elección final para la parte práctica del proyecto se ha decidido especificar de forma más concreta todo el funcionamiento del proceso de estimación.

1. **El Estimador de Dos Pasos**: Propuesto originalmente por Doz, Giannone, y Reichlin (2011), este método surge como una alternativa computacionalmente eficiente a la Máxima Verosimilitud exacta. El procedimiento se descompone en dos fases diferenciadas:

    - *Paso 1 (Consistencia Transversal):* Se estima el espacio generado por los factores mediante PCA sobre el panel de datos. Para ello, habitualmente se trunca el panel original $\mathbf{X}_t$ para obtener un subpanel balanceado o se realiza una estandarización previa. A partir de los factores estimados $\hat{\mathbf{F}}_t$, se obtienen los parámetros del sistema $(\hat{\mathbf{\Lambda}}, \hat{\mathbf{\Phi}}, \hat{\mathbf{\Sigma}}_e, \hat{\mathbf{\Sigma}}_\eta)$ mediante regresiones lineales por Mínimos Cuadrados Ordinarios (MCO).

- *Paso 2 (Eficiencia Dinámica):* Utilizando los parámetros estimados en la primera fase, se proyecta el modelo en una representación de Espacio de Estados. Se aplica el Filtro de Kalman (o el Suavizador) sobre el panel **desbalanceado** original para extraer la estimación final de los factores.

Doz et al. (2011) demuestran que este estimador es consistente cuando $N, T \to \infty$. La principal ventaja es que el Filtro de Kalman actúa como un optimizador de la estimación inicial de PCA, explotando la dinámica autorregresiva de los factores y permitiendo realizar el *nowcasting* sobre el panel con borde dentado de la muestra.

2. **El Algoritmo Quasi-EM (QML)**: Como extensión al método anterior, Doz, Giannone, y Reichlin (2012) proponen el uso del algoritmo de Esperanza-Maximización en un entorno de alta dimensión bajo un enfoque de Máxima Verosimilitud Cuasi-Exacta (*Quasi-Maximum Likelihood*). A diferencia del EM clásico, este enfoque se denomina "Quasi" porque el algoritmo se inicializa mediante el estimador de PCA para garantizar la convergencia hacia el máximo global de la verosimilitud en superficies de alta dimensionalidad.

   La recursión de Doz et al. (2012) demuestra que el algoritmo EM es extremadamente robusto ante especificaciones incorrectas del modelo (v.g., presencia de correlación débil en los errores aleatorios). Matemáticamente, mientras que el estimador de dos pasos realiza una única iteración del filtro de Kalman, el enfoque Quasi-EM itera entre el paso E y el paso M hasta que el incremento de la log-verosimilitud (3.4) sea inferior a un umbral de tolerancia ϵ:

$$\ln L(\boldsymbol{\theta}^{(j+1)}) - \ln L(\boldsymbol{\theta}^{(j)}) < \epsilon.$$

3. **Extensión al algoritmo de Quasi-EM (QML) y Vector de Estado Ampliado**: La consolidación operativa de los métodos híbridos en la macroeconometría de alta dimensión y tiempo real se asienta sobre el trabajo fundamental de Bańbura y Modugno (2014). Los autores generalizaron el algoritmo de Esperanza-Maximización Cuasi-Exacto (Quasi-EM) para superar simultáneamente los tres grandes desafíos estructurales que lastran el *nowcasting* institucional: los patrones arbitrarios de datos ausentes (*ragged edges*), la coexistencia de series de frecuencias mixtas (mensuales y trimestrales) y la persistencia temporal o autocorrelación serial en las perturbaciones.

   Para integrar estas propiedades sin que el sistema pierda la linealidad necesaria para la ejecución del Filtro de Kalman, el modelo experimenta una reestructuración analítica profunda que traslada toda la dinámica estocástica al bloque de transición.

1. Arquitectura del Vector de Estado Ampliado generalizado Sea un panel que contiene N variables observables, desagregadas en N_m indicadores de frecuencia mensual y N_q variables de frecuencia trimestral ($N = N_m + N_q$). Se asume la existencia de r factores comunes que siguen un proceso VAR(p). Para dar cumplimiento al esquema de agregación temporal de flujos y promedios móviles de Mariano y Murasawa (2003), el sistema requiere retener en memoria de forma sistemática un horizonte de 5 meses del factor común $(f_t, f_{t-1}, \ldots, f_{t-4})$.

Asimismo, para purgar la inercia serial de los residuos sin contaminar la señal del ciclo económico, se modelizan las perturbaciones de la ecuación de medida mediante procesos autorregresivos independientes de primer orden, $e_{it} = \rho_i e_{i,t-1} + v_{it}$, donde $v_{it} \sim$ i.i.d. $N(0, \sigma^2_{v,i})$. Debido a la naturaleza de la agregación de las N_q variables trimestrales, sus respectivos errores subyacentes también deben ser acumulados en una estructura de promedios móviles mensuales de 5 periodos.

A partir de estas restricciones, el algoritmo de Bańbura y Modugno (2014) define un **Vector de Estado Ampliado** ($\boldsymbol{\alpha}_t$) de dimensión general $m \times 1$, donde la longitud del espacio latente

queda determinada algebraicamente de forma exacta por la expresión $m = 5r + N_m + 5N_q$:

$$\boldsymbol{\alpha}_t = \begin{pmatrix} \mathbf{f}_t' & \mathbf{f}_{t-1}' & \dots & \mathbf{f}_{t-4}' & \mathbf{e}_{m,t}' & \mathbf{e}_{q,t}' & \mathbf{e}_{q,t-1}' & \dots & \mathbf{e}_{q,t-4}' \end{pmatrix}',$$

donde $\mathbf{f}_t$ representa el bloque de los factores latentes, $\mathbf{e}_{m,t} = (e_{1,t}, \dots, e_{N_m,t})'$ es el bloque que almacena los errores autorregresivos contemporáneos de las series mensuales, y los vectores $\mathbf{e}_{q,t-j}$ componen la Cinta Transportadora Temporal necesaria para la acumulación del ruido de las series trimestrales.

2. Formulación matricial homogénea del sistema Bajo esta representación expandida, el Modelo Factorial Dinámico se formula mediante un sistema homogéneo en el Espacio de Estados libre de errores de medida estocásticos tradicionales:

$$\mathbf{X}_t = \mathbf{C}\boldsymbol{\alpha}_t + \xi_t, \quad \xi_t \sim N(\mathbf{0}, \mathbf{R}).$$

$$\boldsymbol{\alpha}_t = \mathbf{A}\boldsymbol{\alpha}_{t-1} + \mathbf{B}\mathbf{v}_t, \quad \mathbf{v}_t \sim N(\mathbf{0}, \boldsymbol{\Sigma}_v).$$

La **Matriz de Observación o Medida** ($\mathbf{C}$), de dimensión $N \times m$, se particiona en bloques horizontales para diferenciar el tratamiento de la frecuencia mensual y trimestral, forzando los pesos polinómicos de Mariano-Murasawa de forma directa sobre las columnas del estado:

$$\mathbf{C} = \begin{pmatrix} \boldsymbol{\Lambda}_m & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{I}_{N_m} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \frac{1}{3}\boldsymbol{\Lambda}_q & \frac{2}{3}\boldsymbol{\Lambda}_q & \boldsymbol{\Lambda}_q & \frac{2}{3}\boldsymbol{\Lambda}_q & \frac{1}{3}\boldsymbol{\Lambda}_q & \mathbf{0} & \frac{1}{3}\mathbf{I}_{N_q} & \frac{2}{3}\mathbf{I}_{N_q} & \mathbf{I}_{N_q} & \frac{2}{3}\mathbf{I}_{N_q} & \frac{1}{3}\mathbf{I}_{N_q} \end{pmatrix},$$

donde $\boldsymbol{\Lambda}_m$ ($N_m \times r$) y $\boldsymbol{\Lambda}_q$ ($N_q \times r$) representan las matrices de cargas factoriales puras asociadas a las series mensuales y trimestrales, respectivamente.

La **Matriz de Transición del Sistema** ($\mathbf{A}$), de dimensión $m \times m$, gobierna la evolución markoviana mediante una estructura compacta hiper-dispersa que acopla las matrices compañeras del bloque del factor con los coeficientes de persistencia ρ_i:

$$\mathbf{A} = \begin{pmatrix} \boldsymbol{\Phi}_1 & \boldsymbol{\Phi}_2 & \dots & \boldsymbol{\Phi}_p & \dots & \mathbf{0} & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{I}_q & \mathbf{0} & \dots & \mathbf{0} & \dots & \mathbf{0} & \mathbf{0} & \dots & \mathbf{0} \\ \vdots & \ddots & \dots & \vdots & \dots & \vdots & \vdots & \dots & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & \mathbf{0} & \dots & \boldsymbol{\rho}_m & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \dots & \mathbf{0} & \dots & \mathbf{0} & \boldsymbol{\rho}_q & \dots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \dots & \mathbf{0} & \dots & \mathbf{0} & \mathbf{I}_{N_q} & \dots & \mathbf{0} \\ \vdots & \dots & \dots & \vdots & \dots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \dots & \dots & \mathbf{0} & \dots & \mathbf{0} & \mathbf{0} & \dots & \mathbf{0} \end{pmatrix},$$

donde $\boldsymbol{\rho}_m = \text{diag}(\rho_1, \dots, \rho_{N_m})$ y $\boldsymbol{\rho}_q = \text{diag}(\rho_{N_m+1}, \dots, \rho_{N_m+N_q})$. Toda la variabilidad real queda contenida en la matriz de covarianza de la transición $\mathbf{Q} = \mathbf{B}\boldsymbol{\Sigma}_v\mathbf{B}'$, la cual almacena en su diagonal

principal la varianza del choque común junto con las varianzas estructurales $\sigma_{v,i}^2$, rellenando con ceros deterministas las filas vinculadas a las identidades de retraso.

Como consecuencia de este diseño algebraico, el vector de error de medida original colapsa a cero ($\xi_t = \mathbf{0}$). Para evitar la presencia de matrices singulares que invaliden de forma recursiva las inversiones numéricas de la Ganancia de Kalman, se introduce un parámetro de regularización determinista, forzando a la matriz de covarianza de observación a actuar como una identidad escalar infinitesimal:

$$\mathbf{R} = \kappa \mathbf{I}_N, \quad \text{donde} \quad \kappa = 10^{-4}.$$

Esta tolerancia numérica infinitesimal funciona como un amortiguador que estabiliza el filtro sin alterar ni contaminar la señal de la Contabilidad Nacional.

3. El Paso E (Expectation) e Inferencia con Datos Ausentes El algoritmo Quasi-EM opera tratando al Vector de Estado Ampliado $\boldsymbol{\alpha}_t$ como una variable omitida o no observable. En el Paso E de la iteración j-ésima, el objetivo es calcular la esperanza condicional de la función de log-verosimilitud completa basada en el conjunto de información disponible de la muestra completa ($\boldsymbol{\Omega}_T$).

Para gestionar de forma endógena los baches informativos y los bordes dentados del calendario de publicaciones, el Paso E ejecuta un Suavizador de Kalman modificado. Ante la presencia de observaciones faltantes en un instante t, el filtro adapta dinámicamente el sistema evaluando exclusivamente el subespacio de variables disponibles. El suavizador proporciona las esperanzas matemáticas condicionales y los momentos cruzados de segunda etapa:

$$\hat{\boldsymbol{\alpha}}_{t|T} = \mathbb{E}_{\boldsymbol{\theta}^{(j)}}[\boldsymbol{\alpha}_t|\boldsymbol{\Omega}_T], \qquad \mathbf{P}_{t|T} = \mathrm{Var}_{\boldsymbol{\theta}^{(j)}}(\boldsymbol{\alpha}_t|\boldsymbol{\Omega}_T).$$

$$\mathbf{P}_{t,t-1|T} = \mathrm{Cov}_{\boldsymbol{\theta}^{(j)}}(\boldsymbol{\alpha}_t, \boldsymbol{\alpha}_{t-1}|\boldsymbol{\Omega}_T).$$

Para garantizar la convergencia del algoritmo hacia el máximo global de la superficie de verosimilitud y eludir la presencia de óptimos locales, se exige una inicialización paramétrica rigurosa. Las matrices se preestiman mediante Componentes Principales estáticos sobre el subpanel balanceado. Críticamente, la incertidumbre inicial del estado, representada por la matriz de covarianzas $\mathbf{P}_{0|0}$, se deriva analíticamente resolviendo la **Ecuación Discreta de Lyapunov**:

$$\mathbf{P}_{0|0} = \mathbf{A}\mathbf{P}_{0|0}\mathbf{A}' + \mathbf{Q}.$$

La resolución iterativa de esta ecuación lineal garantiza que el Filtro de Kalman inicie su primera recursión asumiendo la varianza incondicional teórica de los procesos estacionarios subyacentes, reflejando fielmente la inercia histórica del panel de indicadores antes de observar la primera realización muestral.

4. El Paso M (Maximization) y la Matriz de Selección Contemporánea En el Paso M, el algoritmo actualiza el vector de parámetros estructurales $\boldsymbol{\theta}^{(j+1)}$ maximizando la función de expectativa calculada en la fase anterior. La gran innovación de Bańbura y Modugno (2014) para posibilitar el cálculo en paneles masivos de alta dimensión consiste en la introducción de una **Matriz de Selección Contemporánea** ($\mathbf{W}_t$).

Se define $\mathbf{W}_t$ como una matriz diagonal de dimensión $N \times N$, cuyos elementos de la diagonal principal adoptan de forma binaria el valor 1 si la variable i es efectivamente observada en el instante temporal t, y el valor 0 en caso de ausencia de información (`NA`). La inclusión analítica de esta estructura altera la obtención de las cargas factoriales, permitiendo que la maximización

se ejecute de forma lineal **fila por fila** mediante regresiones adaptadas de Mínimos Cuadrados Ordinarios (MCO):

$$\mathbf{C}_{i,\cdot}^{(j+1)} = \left(\sum_{t=1}^{T} w_{it} \mathbf{X}_{it} \hat{\boldsymbol{\alpha}}_{t|T}' \right) \left(\sum_{t=1}^{T} w_{it} \left[\hat{\boldsymbol{\alpha}}_{t|T} \hat{\boldsymbol{\alpha}}_{t|T}' + \mathbf{P}_{t|T} \right] \right)^{-1},$$

donde w_{it} es el elemento i-ésimo de la diagonal de $\mathbf{W}_t$, y $\mathbf{C}_{i,\cdot}$ representa la fila i-ésima de la matriz de observaciones expandida. Esta optimización desagregada reduce de forma drástica el coste computacional del algoritmo, transformando un cuello de botella matricial de orden cúbico $O(N^3)$ en una sucesión de operaciones de complejidad puramente lineal $O(N)$ por iteración. El proceso itera entre el Paso E y el Paso M hasta comprobar que el incremento de la log-verosimilitud condicional se sitúa por debajo de un umbral de tolerancia numérica: $\ln L(\boldsymbol{\theta}^{(j+1)}) - \ln L(\boldsymbol{\theta}^{(j)}) < \epsilon$.

5. Post-análisis analítico: Descomposición de Noticias (*News*) Una vez alcanzada la convergencia de los parámetros, esta especificación integrada en el espacio de estados permite formalizar matemáticamente el concepto económico de la Descomposición de Noticias. En el entorno operativo del *nowcasting*, la estimación de la variable de interés (el PIB) se actualiza continuamente ante la llegada de nuevos datos.

Sea $\boldsymbol{\Omega}_{old}$ el conjunto de información disponible inicial y $\boldsymbol{\Omega}_{new}$ el panel expandido tras la publicación de un nuevo grupo de indicadores económicos. La revisión del factor común (o de la proyección del PIB) entre ambos instantes temporales se define de forma exacta como la suma lineal ponderada de las "sorpresas" macroeconómicas contenidas en las nuevas publicaciones, filtradas matemáticamente a través de la Ganancia de Kalman Suavizada:

$$\mathbb{E}[\mathbf{f}_t|\boldsymbol{\Omega}_{new}] - \mathbb{E}[\mathbf{f}_t|\boldsymbol{\Omega}_{old}] = \sum_{i \in \text{News}} \omega_{i,t} \left(\mathbf{X}_{i,v} - \mathbb{E}[\mathbf{X}_{i,v}|\boldsymbol{\Omega}_{old}] \right),$$

donde $\mathbf{X}_{i,v}$ es el dato real publicado para la variable i en el periodo v, el término entre paréntesis representa la innovación económica pura (el dato observado menos la predicción que el modelo tenía de él basándose en la historia pasada), y $\omega_{i,t}$ constituye el vector de ponderación espacial dinámico calculado a partir del multiplicador de ganancia del suavizador. Esta propiedad dota al modelo de una interpretabilidad económica absoluta, garantizando la trazabilidad exacta de cualquier revisión en las perspectivas del crecimiento económico.

3.3.3.4. Enfoque Bayesiano

El **enfoque bayesiano** se ha consolidado como una opción muy utilizada recientemente en el *Nowcasting*. Instituciones de referencia como la **Reserva Federal de Nueva York**, en su *New York Fed Staff Nowcast*, y el **Banco de España**, mediante evoluciones de sus modelos de factores dinámicos, emplean técnicas bayesianas para sortear las limitaciones de la Máxima Verosimilitud convencional en entornos de alta volatilidad y dimensionalidad.

La esencia de este enfoque radica en tratar los parámetros del modelo ($\boldsymbol{\theta}$) no como constantes fijas, sino como variables aleatorias. Mediante la aplicación del **Teorema de Bayes**[5], se combina la información previa del investigador a través de las distribuciones *prior*, con la evidencia contenida en los datos (función de verosimilitud) para obtener la **distribución posterior** de los factores y parámetros:

[5]El Teorema de Bayes se expresa matemáticamente como $P(A|B) = \frac{P(B|A)P(A)}{P(B)}$. En el contexto de la inferencia bayesiana, establece que la probabilidad de que nuestra hipótesis sobre los parámetros sea cierta dados los datos observados (posterior), es proporcional a la probabilidad de observar esos datos si la hipótesis fuera cierta (verosimilitud) multiplicada por nuestra creencia inicial sobre dicha hipótesis (prior).

$$p(\boldsymbol{\theta}, \mathbf{F}_1^T | \mathbf{X}_1^T) \propto p(\mathbf{X}_1^T | \boldsymbol{\theta}, \mathbf{F}_1^T) p(\mathbf{F}_1^T | \boldsymbol{\theta}) p(\boldsymbol{\theta}).$$

Esta metodología ofrece tres modificaciones destacadas para el *nowcasting*:

- **Regularización:** El uso de *priors* informativos actúa como un mecanismo de regularización natural, evitando el sobreajuste que suelen sufrir los modelos MLE cuando el número de series N es elevado en relación a la muestra T.
- **Cuantificación de la Incertidumbre:** A diferencia de los métodos clásicos que suelen utilizar estimaciones puntuales de los factores, el enfoque bayesiano genera distribuciones completas. Esto permite que los intervalos de confianza del PIB reflejen tanto la incertidumbre de los datos faltantes como la incertidumbre sobre los propios parámetros del modelo, permitiendo una cuantificación realista de la incertidumbre real de la estimación que se proporciona.
- **Flexibilidad Estructural:** Facilita la incorporación de características no lineales, como la **volatilidad estocástica** en los errores o parámetros que varían en el tiempo, aspectos fundamentales para capturar cambios de régimen económico como los observados durante la crisis de la COVID-19.

Operativamente, debido a la complejidad de las integrales resultantes, la estimación se realiza mediante métodos de simulación de **Cadenas de Markov Monte Carlo (MCMC)**, destacando el uso del **Muestreador de Gibbs**. Este algoritmo itera sorteando valores de los factores dado los parámetros, y viceversa, convergiendo hacia la distribución conjunta deseada. Es esta capacidad de manejar modelos altamente complejos lo que ha llevado a su adopción en los bancos centrales para la elaboración de proyecciones en tiempo real, además de cuantificar de forma realista la incertidumbre de las estimaciones del modelo, permitiendo cuantificaciones realistas de la probabilidad de una recesión.

3.3.4. Determinación del número de factores

La especificación del número óptimo de factores estáticos (r) y dinámicos (q) es un paso necesario para la estimación consistente del Modelo Factorial Dinámico. Una infraestimación del espacio latente omite dinámicas estructurales relevantes, mientras que una sobreestimación introduce ruido aleatorio en la señal, degradando la eficiencia del pronóstico.

3.3.4.1. Estimación del número de factores estáticos (r)

El número de factores estáticos, r, define la dimensión del subespacio necesario para capturar la varianza transversal del panel. La literatura aborda su estimación mediante enfoques heurísticos y analíticos:

1. Gráfico de sedimentación (*Scree Plot*). El enfoque exploratorio clásico consiste en analizar la distribución de los autovalores (λ_i) de la matriz de covarianzas muestral de $\mathbf{X}_t$, ordenados de forma decreciente. El *scree plot* representa la contribución marginal de cada componente principal a la varianza total. La regla heurística determina r identificando el codo en el gráfico, es decir, el punto donde la caída de los autovalores se estabiliza, indicando que los componentes subsiguientes albergan en gran parte, ruido aleatorio. Un ejemplo de este criterio exploratorio de selección del número de factores se puede observar en la Figura 3.5:

2. Criterios de Información de Bai y Ng (2002) para factores estáticos (r).

La literatura en la que se ha apoyado esta sección asume que la dimensión transversal (N) o la temporal (T) permanecen fijas, lo que invalida el uso de los criterios AIC o BIC convencionales cuando se trabaja con paneles masivos. Bai y Ng (2002) abordan este problema formulando la determinación del número de factores estáticos verdaderos (r) como un problema de selección de modelos para un modelo factorial aproximado, operando bajo el marco asintótico donde $N, T \to \infty$ simultáneamente.

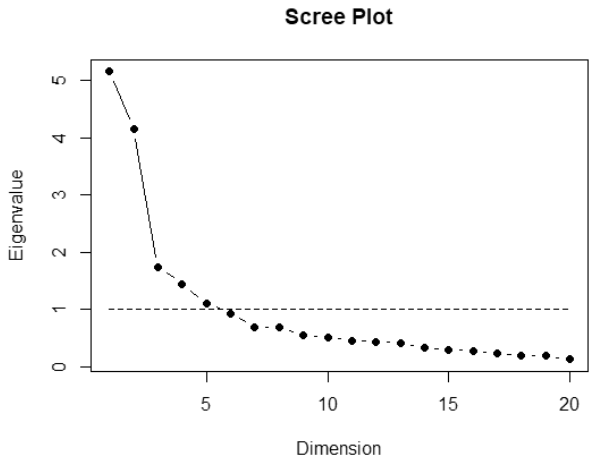


Figura 3.5: Muestra los autovalores de la matriz de covarianzas. La línea punteada horizontal representa el criterio clásico de Kaiser (autovalores > 1), mientras que la caída abrupta o codo sugiere el número óptimo de factores a retener.
.

Para garantizar la estimación consistente del espacio generado por los factores mediante componentes principales, Bai y Ng (2002) imponen los siguientes supuestos sobre la estructura del modelo, permitiendo así una formulación de factor aproximado:

- **Supuesto A (Factores):** $\mathbb{E}\left\|\mathbf{F}_t^0\right\|^4 < \infty$ y $T^{-1}\sum_{t=1}^{T}\mathbf{F}_t^0\mathbf{F}_t^{0\prime} \to \mathbf{\Sigma}_F$ cuando $T \to \infty$ para alguna matriz definida positiva $\mathbf{\Sigma}_F$.

- **Supuesto B (Cargas Factoriales):** $\|\boldsymbol{\lambda}_i\| \leq \bar{\lambda} < \infty$, y $\left\|\frac{\mathbf{\Lambda}^{0\prime}\mathbf{\Lambda}^0}{N} - \mathbf{D}\right\| \to 0$ cuando $N \to \infty$ para alguna matriz $\mathbf{D}$ definida positiva de dimensión $r \times r$.

- **Supuesto C (Dependencia Temporal y Transversal, y Heterocedasticidad):** Existe una constante positiva $M < \infty$, tal que para todo N y T:

  1. $\mathbb{E}(e_{it}) = 0, \quad \mathbb{E}|e_{it}|^8 \leq M.$
  2. $\mathbb{E}\left(\frac{\mathbf{e}_s'\mathbf{e}_t}{N}\right) = \mathbb{E}\left(N^{-1}\sum_{i=1}^{N} e_{is}e_{it}\right) = \gamma_N(s,t)$,
     con $|\gamma_N(s,s)| \leq M$ para todo s, y $T^{-1}\sum_{s=1}^{T}\sum_{t=1}^{T}|\gamma_N(s,t)| \leq M$.
  3. $\mathbb{E}(e_{it}e_{jt}) = \tau_{ij,t}$ con $|\tau_{ij,t}| \leq |\tau_{ij}|$ para algún τ_{ij} y para todo t; además, $N^{-1}\sum_{i=1}^{N}\sum_{j=1}^{N}|\tau_{ij}| \leq M$.
  4. $\mathbb{E}(e_{it}e_{js}) = \tau_{ij,ts}$, y $(NT)^{-1}\sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{t=1}^{T}\sum_{s=1}^{T}|\tau_{ij,ts}| \leq M$.
  5. Para cada par (t,s), $\mathbb{E}\left|N^{-1/2}\sum_{i=1}^{N}[e_{is}e_{it} - \mathbb{E}(e_{is}e_{it})]\right|^4 \leq M$.

- **Supuesto D (Dependencia Débil entre Factores y perturbación aleatoria):**

$$\mathbb{E}\left(\frac{1}{N}\sum_{i=1}^{N}\left\|\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\mathbf{F}_t^0 e_{it}\right\|^2\right) \leq M.$$

Siguiendo con la definición de la suma de residuos al cuadrado promediada obtenida mediante el método de componentes principales asintóticos para k factores como:

$$V(k, \hat{\mathbf{F}}^k) = \min_{\mathbf{\Lambda}, \mathbf{F}^k} \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \left(X_{it} - \boldsymbol{\lambda}_i^{k\prime} \hat{\mathbf{F}}_t^{\ k} \right)^2 .$$

El **Teorema 1** de Bai y Ng (2002) establece que la tasa de convergencia promedio de los factores estimados ($\hat{\mathbf{F}}^k$) hacia el espacio de los factores verdaderos está acotada por $C_{NT}^2 = \min\{N, T\}$.

Teorema 3.2 (Teorema 1 (Bai y Ng, 2002))**.** *Para cualquier $k \geq 1$ fijo, existe una matriz $\mathbf{H}^k$ de dimensión $(r \times k)$ con rango$(\mathbf{H}^k) = \min\{k, r\}$, y definiendo la tasa de convergencia como $C_{NT} = \min\{\sqrt{N}, \sqrt{T}\}$, se cumple que:*

$$C_{NT}^2 \left(\frac{1}{T} \sum_{t=1}^{T} \left\| \hat{\mathbf{F}}_t^k - \mathbf{H}^{k\prime} \mathbf{F}_t^0 \right\|^2 \right) = O_p(1)$$

Lo que se busca es encontrar funciones de penalización, $g(N, T)$, tales que el criterio de la forma:

$$PC(k) = V(k, \hat{\mathbf{F}}^k) + kg(N, T), \tag{3.5}$$

puede estimar consistentemente r. Sea kmax un entero acotado tal que $r <= kmax$.

Los autores demuestran en su **Teorema 2** que la función de penalización $g(N, T)$ que acabamos de representar en la Ecuación 3.5 añadida a $V(k, \hat{\mathbf{F}}^k)$ producirá un estimador asintóticamente consistente del número de factores ($\lim_{N,T\to\infty} P(\hat{k} = r) = 1$) si y solo si cumple las dos condiciones analíticas que postulan:

Teorema 3.3 (Teorema 2 de Bai y Ng, 2002)**.** *Supongamos que se cumplen los Supuestos A-D sobre los factores y los errores, y que los k factores se estiman por el método de componentes principales. Sea $\hat{k} = \arg\min_{0\leq k\leq k_{max}} PC(k)$. Entonces:*

$$\lim_{N,T\to\infty} P(\hat{k} = r) = 1$$

si se cumplen las dos condiciones siguientes cuando $N, T \to \infty$:
(i) $g(N, T) \to 0$
(ii) $C_{NT}^2 \cdot g(N, T) \to \infty$
donde $C_{NT} = \min\{\sqrt{N}, \sqrt{T}\}$.

Bajo estas condiciones, el **Corolario 1** de su trabajo introduce la familia de criterios de información logarítmicos (IC_p), que evitan la necesidad de estimar previamente la varianza de los errores y que estimará de forma consistente el número de factores estáticos, r. La función objetivo a minimizar se define como:

$$IC_p(k) = \ln\left(V(k, \hat{\mathbf{F}}^k) \right) + k \cdot g(N, T).$$

Los 3 criterios propuestos por Bai y Ng (2002) son:

$$PC_{p1}(k) = V(k, \hat{\mathbf{F}}^k) + k\hat{\sigma}^2 \left(\frac{N+T}{NT} \right) \ln\left(\frac{NT}{N+T} \right)$$

$$PC_{p2}(k) = V(k, \hat{\mathbf{F}}^k) + k\hat{\sigma}^2 \left(\frac{N+T}{NT} \right) \ln C_{NT}^2$$

$$PC_{p3}(k) = V(k, \hat{\mathbf{F}}^k) + k\hat{\sigma}^2 \left(\frac{\ln C_{NT}^2}{C_{NT}^2} \right),$$

donde $\hat{\sigma}^2$ es un estimador consistente de $(NT)^{-1} \sum_{i=1}^{N} \sum_{t=1}^{T} \mathbb{E}(e_{it})^2$.

El Corolario 1 de su investigación lleva a considerar también los siguientes criterios logarítmicos, que evitan la dependencia del estimador de la varianza $\hat{\sigma}^2$:

$$IC_{p1}(k) = \ln\left(V(k, \hat{\mathbf{F}}^k)\right) + k\left(\frac{N+T}{NT}\right)\ln\left(\frac{NT}{N+T}\right)$$
$$IC_{p2}(k) = \ln\left(V(k, \hat{\mathbf{F}}^k)\right) + k\left(\frac{N+T}{NT}\right)\ln C_{NT}^2$$
$$IC_{p3}(k) = \ln\left(V(k, \hat{\mathbf{F}}^k)\right) + k\left(\frac{\ln C_{NT}^2}{C_{NT}^2}\right)$$

3. Estimadores *Eigenvalue Ratio* y *Growth Ratio* (Ahn y Horenstein, 2013). Los criterios de Bai y Ng (2002) tienen el defecto de verse demasiado alterados ante la elección arbitraria del número máximo de factores (k_{max}) y tienden a sobreestimar r en presencia de correlación cruzada residual. Para solventar estas deficiencias, Ahn y Horenstein (2013) introdujeron dos estimadores que no dependen de hiperparámetros ni de la varianza del error. El estimador *Eigenvalue Ratio* (ER) maximiza el ratio de autovalores adyacentes:

$$\hat{r}_{ER} = \underset{1\leq k\leq k_{max}}{\text{máx}} \frac{\lambda_k}{\lambda_{k+1}}.$$

Analíticamente, este estimador formaliza la búsqueda del precipicio relativo máximo en el gráfico de sedimentación. Al depender de ratios de autovalores, este método aísla eficientemente la señal común del ruido aleatorio fuertemente correlacionado.

3.3.4.2. Estimación del número de factores dinámicos (q)

Mientras que r define la dimensión del subespacio abarcado por los factores estáticos (que capturan la dinámica acumulada), q representa el número de *shocks* primitivos o factores dinámicos que impulsan las fluctuaciones comunes del sistema ($q \leq r$) (Stock y Watson, 2016). Formalmente, se asume que los factores estáticos $\mathbf{F}_t$ siguen un proceso VAR de orden p dado por $\mathbf{\Phi}(L)\mathbf{F}_t = \mathbf{u}_t$, donde las innovaciones $\mathbf{u}_t$ están conducidas por un vector de dimensión reducida (Stock y Watson, 2016). Decimos que el sistema está impulsado por q *shocks* si existe una matriz $\mathbf{G}$ de dimensión $r \times q$ con rango q tal que:

$$\mathbf{u}_t = \mathbf{G}\boldsymbol{\eta}_t,$$

donde $\boldsymbol{\eta}_t$ es un vector $q \times 1$ de *shocks* dinámicos mutuamente incorrelados (Stock y Watson, 2005, 2016). Bajo esta estructura, la matriz de covarianzas de las innovaciones del vector de estado, definida como $\mathbf{\Sigma}_u = \mathbb{E}(\mathbf{u}_t\mathbf{u}_t') = \mathbf{G}\mathbf{\Sigma}_\eta\mathbf{G}'$, posee un rango teórico exactamente igual a q (Bai y Ng, 2007; Stock y Watson, 2005, 2016).

Para determinar q en entornos de alta dimensionalidad, Bai y Ng (2007) proponen un procedimiento en el dominio temporal basado en la descomposición espectral de la matriz de varianzas-covarianzas de los residuos del VAR estimado sobre los factores, $\hat{\mathbf{\Sigma}}_u$. Dado que los factores son latentes, el procedimiento se aplica sobre los estimadores de componentes principales $\hat{\mathbf{F}}_t$ Bai y Ng (2007); Krantz y Bagdziunas (2023).

El **Lema 2** de Bai y Ng (2007) garantiza que la matriz de covarianzas muestral de los residuos del VAR, $\hat{\mathbf{\Sigma}}_u = \frac{1}{T}\sum_{t=1}^{T}\hat{\mathbf{u}}_t\hat{\mathbf{u}}_t'$, converge a una rotación de la verdadera matriz de covarianzas a una tasa determinada por la dimensión más restrictiva del panel:

$$\text{mín}[\sqrt{N}, \sqrt{T}]\left(\hat{\mathbf{\Sigma}}_u - \mathbf{H}^*\mathbf{\Sigma}_u\mathbf{H}^{*\prime}\right) = O_p(1).$$

El test evalúa cuántos autovalores de $\hat{\mathbf{\Sigma}}_u$ (o de la matriz de correlaciones $\hat{\mathbf{S}}_u$) son asintóticamente distintos de cero. Sean $c_1 > c_2 \geq \cdots \geq c_r \geq 0$ los autovalores ordenados. Se definen la k-ésima normalización del k-ésimo autovalor, las medidas de desviación marginal del (k+1)-ésimo autovalor

y de contribución acumulada de los autovalores respecto a la hipótesis nula de rango($\hat{\mathbf{\Sigma}}_u$) = q y que $c_k = 0$ para $k > q$, respectivamente:

$$\hat{D}_k = \left(\frac{c_k^2}{\sum_{j=1}^{r} c_j^2} \right)^{1/2}, \qquad \hat{D}_{1,k} = \left(\frac{c_{k+1}^2}{\sum_{j=1}^{r} c_j^2} \right)^{1/2}, \qquad \hat{D}_{2,k} = \left(\frac{\sum_{j=k+1}^{r} c_j^2}{\sum_{j=1}^{r} c_j^2} \right)^{1/2}.$$

La **Proposición 2** de Bai y Ng (2007) establece estimadores consistentes para q. Para constantes $m > 0$ y $\delta \in (0, 1/2)$, se definen los conjuntos que satisfacen una cota de tolerancia menguante:

$$\mathcal{K}_3 = \left\{ k : \hat{D}_{1,k} < \frac{m}{C_{NT}^{\delta}} \right\}$$
$$\mathcal{K}_4 = \left\{ k : \hat{D}_{2,k} < \frac{m}{C_{NT}^{\delta}} \right\}$$

donde $C_{NT} = \text{mín}[\sqrt{N}, \sqrt{T}]$. Los estimadores se definen como $\hat{q}_3 = \text{mín}\{k \in \mathcal{K}_3\}$ y $\hat{q}_4 = \text{mín}\{k \in \mathcal{K}_4\}$. Bai y Ng (2007) demuestran que los estimadores convergen en probabilidad al verdadero valor de *shocks primitivos*, $\hat{q}_3, \hat{q}_4 \xrightarrow{p} q$ cuando $N, T \to \infty$.

En la práctica, se prefiere el uso de la matriz de correlaciones $\hat{\mathbf{S}}_u$ para asegurar invarianza de escala, empleando comúnmente $m = 1$. Como alternativa robusta, el método de Amengual y Watson (2007) permite estimar q aplicando los criterios de información de Bai y Ng (2002) sobre los residuos de la proyección de X_t frente a los retardos de los factores estáticos.

La determinación del número de factores dinámicos depende de la dimensión estática previa (r). Siguiendo a Stock y Watson (2016), se presenta una comparativa de los estimadores obtenidos para el panel de datos, donde se evidencia la divergencia entre los criterios basados en el rango de innovaciones y los basados en información residual.

- Bai y Ng (2007): Se centra en el análisis de los autovalores de la matriz de covarianza de las innovaciones del VAR del factor, $\mathbf{\Sigma}_u$. El objetivo es identificar el número de valores propios que son asintóticamente distintos de cero bajo la premisa de singularidad dinámica.

- Amengual y Watson (2007): Propone aplicar los criterios de información IC_{p2} de Bai y Ng (2002) sobre los residuos de una regresión de X_t proyectada sobre los retardos de los factores estáticos. Este método es a menudo más robusto cuando la señal de los choques dinámicos es débil en comparación con el ruido aleatorio.

Como se observa en la Tabla 3.1, el estimador de Bai y Ng (2007) tiende a ser menos conservador en la identificación de shocks, mientras que Amengual y Watson (2007) suele favorecer una mayor parsimonia dinámica. Esta diferencia radica en que el segundo método penaliza más estrictamente la inclusión de innovaciones que no aportan un incremento significativo al R^2 de la traza del panel X_t.

En resumen, el orden lógico del procedimiento es:

1. Estimar el número de factores estáticos r mediante criterios de información IC_p Bai y Ng (2002) o el estimador *Eigenvalue Ratio* de Ahn y Horenstein (2013).

2. Ajustar un modelo VAR(p) sobre $\hat{\mathbf{F}}_t$, seleccionando el lag p mediante el criterio BIC Stock y Watson (2016).

3. Aplicar los criterios de Bai y Ng (2007) sobre los autovalores de la matriz de covarianzas de los residuos del VAR para identificar el rango q.

Cabe destacar que las aplicaciones reales del modelo se tiene a favorecer una especificación parsimoniosa (frecuentemente $q = 1$ o $q = 2$) para facilitar la interpretación del factor como un indicador del ciclo económico general (Mariano y Murasawa, 2003; Stock y Watson, 2016).

Tabla 3.1: Comparativa de estimadores para q ante diferentes niveles de r
.

Factores Estáticos (r)	Bai-Ng (2007) ($\hat{q}_{BN}$)	Amengual-Watson (2007) ($\hat{q}_{AW}$)	Decisión Empírica
$r = 2$	2	1	$q = 1$
$r = 4$	3	2	$q = 2$
$r = 8$	5	3	$q = 3$
$r = r^*$ (Óptimo IC_{p2})	4	3	**Sugerido:** $q = 3$

Nota: Los valores son ilustrativos de la tendencia observada en la literatura de grandes paneles donde $r > q$ debido a la estructura de retardos de los factores.

3.3.5. Inferencia y propiedades asintóticas de los estimadores

El análisis riguroso de los MFD exige establecer las condiciones bajo las cuales los métodos de estimación descritos en la sección anterior proporcionan resultados estadísticamente válidos. En el entorno clásico de series temporales, donde la dimensión transversal (N) es pequeña y fija mientras que la dimensión temporal (T) $\to \infty$, las propiedades de los estimadores de Máxima Verosimilitud (ML) están bien documentadas, alcanzando la consistencia y la eficiencia asintótica tradicional.

Sin embargo, en el contexto del *nowcasting* en el que se ha ido presentando durante el presente proyecto se opera en entornos de alta dimensión, donde tanto $N \to \infty$ como $T \to \infty$. En este contexto, el análisis de las propiedades de los estimadores de los factores latentes ($\hat{\mathbf{F}}_t$) y de sus cargas asociadas ($\hat{\mathbf{\Lambda}}$) resulta más complejo, siendo fundamental distinguir entre el enfoque no paramétrico (PCA) y los enfoques basados en verosimilitud cuasi-exacta (QML).

3.3.5.1. Propiedades del estimador de Componentes Principales (PCA)

La teoría inferencial para el estimador de Componentes Principales (PCA) en entornos de gran dimensión se formaliza en Bai (2003). Este marco analítico asume la estructura del **Modelo Factorial Aproximado** (Chamberlain y Rothschild, 1983), la cual admite la presencia de correlación débil, tanto transversal como serial, así como heterocedasticidad en los errores aleatorios (e_{it}).

Dado que los factores ($\mathbf{F}_t$) y las cargas factoriales ($\boldsymbol{\lambda}_i$) no son identificables por separado, el estimador de ACP identifica el espacio generado por los factores salvo una matriz de transformación por rotación invertible $\mathbf{H}$. En consecuencia, el estimador $\tilde{\mathbf{F}}_t$ converge a $\mathbf{H}'\mathbf{F}_t$, y $\tilde{\boldsymbol{\lambda}}_i$ converge a $\mathbf{H}^{-1}\boldsymbol{\lambda}_i$. Bajo el supuesto de doble asintótica ($N, T \to \infty$), el comportamiento estadístico del sistema queda definido por las siguientes distribuciones límite y tasas de convergencia:

1. Distribución asintótica y convergencia uniforme de los factores estimados: Bajo la condición de que el crecimiento de la muestra temporal domine al de la sección cruzada tal que $\frac{\sqrt{N}}{T} \to 0$, el estimador del factor en el instante t sigue una distribución Normal asintótica:

$$\sqrt{N}(\tilde{\mathbf{F}}_t - \mathbf{H}'\mathbf{F}_t) \xrightarrow{d} \mathcal{N}(0, \mathbf{V}^{-1}\mathbf{Q}\mathbf{\Gamma}_t\mathbf{Q}'\mathbf{V}^{-1}).$$

donde $\mathbf{V}$ es la matriz diagonal de los r mayores autovalores de la matriz de covarianza muestral, $\mathbf{Q}$ denota la matriz límite de rotación asintótica, y $\mathbf{\Gamma}_t = \lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{N} \boldsymbol{\lambda}_i \boldsymbol{\lambda}_j' \mathbb{E}(e_{it}e_{jt})$ modeliza la estructura de covarianza transversal de los errores. La tasa de convergencia asintótica para $\tilde{\mathbf{F}}_t$

es mín$\{\sqrt{N}, T\}$. Adicionalmente, el estimador presenta convergencia uniforme en toda la muestra, acotándose la desviación máxima escalar mediante máx$_t \|\tilde{\mathbf{F}}_t - \mathbf{H}'\mathbf{F}_t\| = \mathcal{O}_p(T^{-1/2}) + \mathcal{O}_p((T/N)^{1/2})$.

2. Distribución asintótica de las cargas factoriales: Si el crecimiento de la dimensión transversal domina al temporal estableciendo que $\frac{\sqrt{T}}{N} \to 0$, el estimador de las cargas factoriales converge en distribución según:

$$\sqrt{T}(\tilde{\boldsymbol{\lambda}}_i - \mathbf{H}^{-1}\boldsymbol{\lambda}_i) \xrightarrow{d} \mathcal{N}\left(0, (\mathbf{Q}')^{-1}\boldsymbol{\Phi}_i\mathbf{Q}^{-1}\right),$$

siendo $\boldsymbol{\Phi}_i = \text{plim}_{T\to\infty} \frac{1}{T}\sum_{s=1}^{T}\sum_{t=1}^{T} \mathbb{E}[\mathbf{F}_t\mathbf{F}'_s e_{is}e_{it}]$ la matriz que captura la autocorrelación temporal del término de error aleatorio para la variable i. En este caso, la tasa de convergencia es mín$\{\sqrt{T}, N\}$.

3. Distribución asintótica de los componentes comunes: El componente común $C_{it} = \boldsymbol{\lambda}_i'\mathbf{F}_t$ resulta directamente identificable en la estimación dado que $\tilde{C}_{it} = \tilde{\boldsymbol{\lambda}}_i'\tilde{\mathbf{F}}_t = \boldsymbol{\lambda}_i'(\mathbf{H}^{-1})'\mathbf{H}'\mathbf{F}_t = \boldsymbol{\lambda}_i'\mathbf{F}_t$. El estimador del componente común converge a su valor poblacional independientemente de la trayectoria relativa de crecimiento entre N y T:

$$\left(\frac{1}{N}V_{it} + \frac{1}{T}W_{it}\right)^{-1/2} (\tilde{C}_{it} - C_{it}) \xrightarrow{d} \mathcal{N}(0,1),$$

con $V_{it} = \boldsymbol{\lambda}_i'\boldsymbol{\Sigma}_\Lambda^{-1}\boldsymbol{\Gamma}_t\boldsymbol{\Sigma}_\Lambda^{-1}\boldsymbol{\lambda}_i$ y $W_{it} = \mathbf{F}_t'\boldsymbol{\Sigma}_F^{-1}\boldsymbol{\Phi}_i\boldsymbol{\Sigma}_F^{-1}\mathbf{F}_t$. El límite se alcanza a una tasa de $\delta_{NT} =$ mín$\{\sqrt{N}, \sqrt{T}\}$.

4. Consistencia uniforme y condiciones bajo dimensión temporal finita: En un supuesto analítico alternativo donde la dimensión temporal es finita (T fijo y $N \to \infty$), Bai (2003) demuestra que la consistencia del estimador $\tilde{\mathbf{F}}_t$ exige como condición matemática necesaria y suficiente el cumplimiento de la ortogonalidad asintótica transversal (lím$_{N\to\infty} \frac{1}{N}\sum_{i=1}^{N} e_{it}e_{is} = 0$ para $t \neq s$) y la homocedasticidad cruzada (lím$_{N\to\infty} \frac{1}{N}\sum_{i=1}^{N} e_{it}^2 = \sigma^2$). El incumplimiento empírico de estas condiciones justifica la exigencia matemática de la doble asintótica ($N, T \to \infty$) para modelizar paneles dinámicos complejos.

Finalmente, el resultado de consistencia uniforme establece que el error de estimación del factor tiende a cero en probabilidad a un ritmo que iguala o supera la tasa estándar de convergencia de la muestra (mín$\{\sqrt{N}, \sqrt{T}\}$). Esta velocidad de convergencia es estadísticamente suficiente para garantizar que el uso del factor estimado $\tilde{\mathbf{F}}_t$, en sustitución del componente latente no observable $\mathbf{F}_t$ en regresiones de segunda etapa, no introduzca sesgos por error de medición ni contamine las distribuciones límite de los parámetros estructurales estimados posteriormente. Esta propiedad analítica inhibe el sesgo asintótico por error de medición, conocido en la literatura econométrica como el problema de regresores generados[6].

3.3.5.2. Propiedades de los estimadores híbridos y Máxima Verosimilitud en alta dimensión

Si bien el PCA ofrece garantías asintóticas robustas, su ineficiencia en muestras finitas bajo dinámicas complejas y su incapacidad para tratar bordes dentados en el panel de datos de forma endógena

[6]El problema de los regresores generados, formalizado metodológicamente por Pagan (1984) y Murphy y Topel (1985), se materializa cuando una variable explicativa utilizada en un modelo es el resultado de una estimación previa. La omisión de la varianza estadística inherente a dicha estimación inicial provoca una subestimación de los errores estándar en la ecuación estructural, distorsionando la inferencia y produciendo contrastes de hipótesis espurios. En el marco factorial, dado que $\|\tilde{\mathbf{F}}_t - \mathbf{H}'\mathbf{F}_t\| \xrightarrow{p} 0$ a tasas consistentes bajo la doble asintótica paramétrica, el efecto del error de aproximación sobre las varianzas de segunda etapa diverge hacia cero, legitimando el tratamiento de los factores como variables observadas con certidumbre, supuesto base para la construcción de modelos FAVAR (Bernanke y Boivin, 2003) y ecuaciones predictivas directas.

justifican el uso de métodos basados en el Filtro de Kalman. La inferencia en este dominio de alta dimensión ha sido formalizada analíticamente por Doz et al. (2011, 2012).

Para el **Estimador de Dos Pasos** (*Two-Step Estimator*), Doz et al. (2011) demuestran que la aplicación del Suavizador de Kalman sobre los parámetros extraídos inicialmente mediante PCA produce estimaciones de los factores que son consistentes. De forma análoga a los resultados de Bai (2003), bajo supuestos de MFD aproximado, la estimación del factor mediante el Filtro de Kalman converge al espacio verdadero a medida que $N, T \to \infty$. La ventaja cualitativa de este estimador no reside en una mejor tasa de convergencia asintótica -que sigue siendo gobernada por mín$\{\sqrt{N}, \sqrt{T}\}$, sino en su mayor eficiencia en muestras finitas, al incorporar explícitamente la estructura autorregresiva (VAR) de los factores en la estimación del estado.

Por su parte, el **Estimador Quasi-Máxima Verosimilitud (QML)** optimizado vía el algoritmo EM (Doz et al., 2012) extiende estas propiedades. Los autores demuestran que, incluso si el MFD se especifica erróneamente asumiendo que la matriz de errores aleatorios $\mathbf{\Sigma}_e$ es estrictamente diagonal (MFD exacto), el maximizador de esta función de cuasi-verosimilitud sigue produciendo estimadores consistentes para los factores en el límite de alta dimensión ($N \to \infty$).

La distribución asintótica bajo el enfoque QML garantiza que:

$$\hat{\mathbf{F}}_t^{QML} - \mathbf{H}'\mathbf{F}_t \xrightarrow{p} 0 \quad \text{cuando} \quad N, T \to \infty.$$

Por ello en la literatura especializada tienden defender una especificación paramétrica deliberadamente simplificada (MFD exacto) dentro del algoritmo EM para evitar la inestabilidad de la matriz Hessiana y la explosión dimensional, sabiendo que el error de especificación en la estructura transversal de los residuos tiende a desvanecerse asintóticamente gracias al gran tamaño del panel de indicadores (N).

3.3.5.3. Estimación de covarianzas e intervalos de confianza

Los teoremas de distribución asintótica proporcionan el marco analítico para realizar inferencia sobre los factores y los componentes comunes. No obstante, las matrices de covarianza de las distribuciones límite están expresadas en función de los parámetros poblacionales verdaderos ($\mathbf{F}^0, \mathbf{\Lambda}^0, \mathbf{\Sigma}_e$), los cuales son inobservables. Para operacionalizar la inferencia, Bai (2003) demuestra que la sustitución de estos parámetros poblacionales por sus correspondientes estimadores de Componentes Principales ($\tilde{\mathbf{F}}, \tilde{\mathbf{\Lambda}}$) genera matrices de covarianza estimadas que convergen consistentemente a las poblacionales.

1. Estimación de las matrices de covarianza: Bajo el supuesto de independencia en la sección cruzada para la construcción de la varianza del factor, los estimadores consistentes se definen algebraicamente de la siguiente forma:

- **Varianza de los factores estimados ($\hat{\mathbf{\Pi}}_t$):** Dado que $\frac{\tilde{\mathbf{F}}'\tilde{\mathbf{F}}}{T} = \mathbf{I}_r$, el estimador consistente de la varianza asintótica del factor en t se reduce a:

$$\hat{\mathbf{\Pi}}_t = \mathbf{V}_{NT}^{-1} \left(\frac{1}{N} \sum_{i=1}^{N} \hat{e}_{it}^2 \tilde{\boldsymbol{\lambda}}_i \tilde{\boldsymbol{\lambda}}_i' \right) \mathbf{V}_{NT}^{-1},$$

  donde $\mathbf{V}_{NT}$ es la matriz diagonal de los r mayores autovalores de $\frac{1}{NT}\mathbf{X}\mathbf{X}'$, y $\hat{e}_{it} = X_{it} - \tilde{\boldsymbol{\lambda}}_i'\tilde{\mathbf{F}}_t$.

- **Varianza de las cargas factoriales ($\hat{\mathbf{\Theta}}_i$):** Se estima mediante un estimador HAC (como Newey-West) [7] aplicado sobre los residuos $\tilde{\mathbf{F}}_t\hat{e}_{it}$ para aproximar consistentemente la varianza teórica límite definida como:

$$\mathbf{\Theta}_i = (\mathbf{Q}')^{-1}\mathbf{\Omega}_i\mathbf{Q}^{-1}.$$

[7] Los estimadores HAC (*Heteroskedasticity and Autocorrelation Consistent*), popularizados por Newey y West (1987), permiten estimar de forma consistente la matriz de varianzas y covarianzas asintótica cuando los términos de error violan los supuestos clásicos de Gauss-Markov (esfera de errores). En el contexto del Modelo Factorial Aproximado, donde se admite explícitamente que los errores aleatorios e_{it} pueden presentar heterocedasticidad y una correlación serial débil, la varianza asintótica no depende únicamente de la varianza contemporánea, sino de la suma infinita de todas las

- **Varianza de los componentes comunes ($\hat{V}_{it}$ y $\hat{W}_{it}$):** El error de estimación del componente común se parametriza mediante las contrapartidas muestrales de las formas cuadráticas asintóticas:

$$\hat{V}_{it} = \tilde{\boldsymbol{\lambda}}_i' \left(\frac{\tilde{\boldsymbol{\Lambda}}'\tilde{\boldsymbol{\Lambda}}}{N} \right)^{-1} \left(\frac{1}{N} \sum_{k=1}^{N} \hat{e}_{kt}^2 \tilde{\boldsymbol{\lambda}}_k \tilde{\boldsymbol{\lambda}}_k' \right) \left(\frac{\tilde{\boldsymbol{\Lambda}}'\tilde{\boldsymbol{\Lambda}}}{N} \right)^{-1} \tilde{\boldsymbol{\lambda}}_i$$

.

$$\hat{W}_{it} = \tilde{\mathbf{F}}_t' \hat{\boldsymbol{\Theta}}_i \tilde{\mathbf{F}}_t$$

.

A partir de estos estimadores, es factible construir intervalos de confianza asintóticos. Siguiendo específicamente los teoremas 3 y 6 de Bai (2003), el intervalo al $(1-\alpha)\,\%$ para el componente común poblacional C_{it} se calcula como:

$$IC_{1-\alpha}(C_{it}) = C_{it} \pm z_{\alpha/2} \left(\frac{1}{N}\hat{V}_{it} + \frac{1}{T}\hat{W}_{it} \right)^{1/2}$$

.

2. Inferencia en muestras finitas: Métodos de remuestreo (Bootstrap): La construcción de los intervalos de confianza asintóticos descritos exige el cumplimiento del marco bidimensional infinito $(N, T \to \infty)$. En aplicaciones macroeconómicas donde las dimensions de la muestra están rígidamente acotadas, la aproximación asintótica Normal tiende a infraestimar la verdadera variabilidad de los parámetros, generando intervalos inferiores al nivel nominal estipulado.

Para asegurar la validez de la inferencia en muestras estrictamente finitas, se recurre a técnicas de remuestreo basadas en algoritmos de *Bootstrap*. En el marco de los modelos factoriales, el procedimiento operativo estandarizado sigue los siguientes pasos algorítmicos:

1. *Estimación inicial:* Se estiman los factores $\tilde{\mathbf{F}}_t$, las cargas $\tilde{\boldsymbol{\Lambda}}$ y los residuos muestrales $\hat{e}_{it}$ sobre la muestra original.

2. *Remuestreo de residuos:* Para conservar la estructura temporal y transversal presente en el Modelo Factorial Aproximado, se aplican técnicas de remuestreo por bloques o remuestreo salvaje sobre los vectores de residuos $\hat{\mathbf{e}}_t$. Esto genera series de perturbaciones pseudo-aleatorias $\hat{e}_{it}^{(b)}$.

3. *Generación de pseudo-paneles:* Se construye el panel de datos simulado sumando la señal original estimada y el ruido aleatorio remuestreado: $X_{it}^{(b)} = \tilde{\boldsymbol{\lambda}}_i' \tilde{\mathbf{F}}_t + \hat{e}_{it}^{(b)}$.

4. *Re-estimación:* Se aplica el estimador (PCA o Algoritmo EM) sobre el nuevo panel $\mathbf{X}^{(b)}$ para obtener $\tilde{\mathbf{F}}_t^{(b)}$ y $\tilde{\boldsymbol{\Lambda}}^{(b)}$.

5. *Construcción de intervalos empíricos:* El proceso se itera B veces (típicamente $B \geq 1000$). Los límites del intervalo de confianza para $\tilde{\mathbf{F}}_t$ se determinan extrayendo directamente los percentiles $\alpha/2$ y $1-\alpha/2$ de la distribución empírica generada por las B réplicas estimadas.

Este enfoque integra la incertidumbre derivada tanto de la extracción de la señal como de la estimación de los parámetros de ponderación espacial, proporcionando bandas de error precisas independientemente del tamaño muestral de N o T.

---

autocovarianzas rezagadas. El estimador de Newey-West aproxima esta suma utilizando los residuos empíricos del PCA ($\hat{e}_{it}$) y aplicando un esquema de ponderación decreciente sobre los retardos. Esta ponderación garantiza analíticamente que la matriz de covarianza resultante sea semidefinida positiva, condición *sine qua non* para la validez de la inferencia estadística y la construcción de intervalos de confianza.

3.3.6. Predicción y Nowcasting

Una vez garantizada la consistencia teórica de los parámetros estructurales del sistema ($\hat{\boldsymbol{\theta}}$) e introducido el algoritmo seleccionado para la extracción de las variables latentes, el objetivo principal del modelo se desplaza hacia la estimación de la variable de interés en el instante actual ($h = 0$). Dado que las fuerzas sectoriales comunes se extraen en frecuencia mensual y la variable de interés se observa únicamente a nivel trimestral, el ejercicio predictivo enfrenta un problema de alineación temporal y asincronía distributiva. Para resolver esta distorsión, la literatura macroeconómica ofrece dos enfoques metodológicos principales:

3.3.6.1. Enfoque Secuencial: Agregación temporal y Ecuaciones Puente (*Bridge Equations*)

Este primer paradigma conceptual consiste en disociar formalmente la etapa de extracción de la señal macroeconómica de la fase de predicción final del crecimiento (Baffigi, Golinelli, y Parigi, 2004). El procedimiento se ejecuta de forma secuencial en dos fases diferenciadas:

En una primera etapa, el Filtro de Kalman (o el modelo VAR subyacente que gobierna el vector latente) proyecta de forma iterativa los factores mensuales hacia aquellos horizontes no observados en el borde dentado de la muestra mediante su propia inercia autorregresiva ($\hat{\mathbf{f}}_{t+h|t} = \mathbf{A}^h \hat{\mathbf{f}}_{t|t}$). Seguidamente, el vector de factores pronosticados en frecuencia mensual se transforma a su contrapartida trimestral equivalente. Siguiendo la aproximación lineal de Mariano y Murasawa (2003), la tasa de crecimiento trimestral se modeliza como un promedio móvil ponderado de las tasas intermensuales. El factor trimestralizado resultante, denotado como $\mathbf{f}_t^Q$, queda determinado de forma unívoca a través de la aplicación de los pesos fijos $(1/3, 2/3, 1, 2/3, 1/3)$:

$$\mathbf{f}_t^Q \approx \frac{1}{3}\mathbf{f}_t + \frac{2}{3}\mathbf{f}_{t-1} + \mathbf{f}_{t-2} + \frac{2}{3}\mathbf{f}_{t-3} + \frac{1}{3}\mathbf{f}_{t-4}$$

.

En una segunda etapa, la estimación del PIB se ejecuta proyectando la variable real observada sobre el espacio generado por el factor trimestralizado mediante una **Ecuación Puente** auxiliar estimada por Mínimos Cuadrados Ordinarios (MCO):

$$\hat{y}_t^Q = \alpha + \boldsymbol{\beta}' \mathbf{f}_t^Q + \varepsilon_t$$

.

En el marco de la econometría clásica, utilizar un regresor previamente estimado ($\mathbf{f}_t^Q$) en una regresión de segunda etapa introduciría un problema de error de medida, provocando un sesgo de atenuación sobre los coeficientes. Sin embargo, en el entorno de los Modelos Factoriales Dinámicos, este procedimiento queda rigurosamente validado bajo el supuesto de doble asintótica. Gracias a la propiedad de **consistencia uniforme** demostrada por Bai (2003), el error de aproximación del factor tiende a cero a una tasa de mín$\{\sqrt{N}, T\}$. Consecuentemente, al trabajar con paneles de gran escala ($N \to \infty$), la varianza del error de estimación de las variables latentes se desvanece asintóticamente.

3.3.6.2. Enfoque Integrado: Inserción directa en el Espacio de Estados

El segundo enfoque introduce la variable trimestral objetivo de forma directa dentro del panel general de indicadores observables $\mathbf{X}_t$, tal como se formuló analíticamente en la Sección 3.3.3.3 bajo la doctrina de Bańbura y Modugno (2014).

Bajo esta especificación, el PIB se presenta como una serie mensual continua afectada por un patrón de datos ausentes, donde los datos reales trimestrales son únicamente visibles en el tercer mes de cada trimestre (marzo, junio, septiembre y diciembre) y se codifican como `NA` en los meses intermedios. Al estar las restricciones algebraicas de agregación de Mariano y Murasawa (2003) inyectadas de forma endógena dentro de la matriz de observación $\mathbf{C}$ y la matriz compañera de transición $\mathbf{A}$, el Filtro de Kalman opera la predicción de manera unificada.

La principal ventaja de este enfoque integrado es su eficiencia: el algoritmo procesa el PIB como una variable mensualizada expuesta al *ragged edge*. Durante aquellos meses donde no se dispone de publicación oficial, la rutina del filtro omite la fase de corrección y computa automáticamente el *nowcast* proyectando el estado latente a partir de la covarianza contemporánea extraída del resto de los indicadores mensuales disponibles, explotando la totalidad de la información cruzada en un único paso estocástico.

3.3.6.3. Post-análisis: Evaluación pseudo-tiempo real y Descomposición de Noticias

Una vez que el sistema dinámico genera las proyecciones del ciclo, se configuran dos procedimientos analíticos de validación y diagnóstico fundamentales para evaluar la consistencia del modelo:

1. Evaluación de precisión fuera de la muestra: Para medir la precisión del *nowcasting* se simula el flujo asíncrono de información que el investigador enfrentaría en tiempo real. Para ello, se ejecutan ejercicios predictivos recursivos en **pseudo-tiempo real**, truncando la muestra retrospectivamente y expandiendo la ventana de información mes a mes.

Para cuantificar la magnitud del error cometido en un horizonte de evaluación fuera de la muestra de tamaño H, se definen matemáticamente dos indicadores de referencia: la Raíz del Error Cuadrático Medio ($RMSE$) y el Error Absoluto Medio (MAE):

$$RMSE = \sqrt{\frac{1}{H}\sum_{h=1}^{H}\left(y_{t+h} - \hat{y}_{t+h|t}\right)^2}$$

.

$$MAE = \frac{1}{H}\sum_{h=1}^{H}\left|y_{t+h} - \hat{y}_{t+h|t}\right|$$

.

El $RMSE$ aplica una ponderación cuadrática sobre los residuos, penalizando de forma desproporcionada la presencia de desviaciones extremas o errores de cola, mientras que el MAE ofrece una métrica de penalización estrictamente lineal, aportando una mayor robustez frente a valores atípicos transitorios. En la fase de post-estimación, el interés analítico del modelo no se reduce a la obtención de la cifra puntual del crecimiento esperado, sino a garantizar la interpretabilidad de sus variaciones. A medida que transcurre el trimestre económico, la publicación de nuevos datos expande el conjunto de información disponible (de $\mathbf{\Omega}_{old}$ a $\mathbf{\Omega}_{new}$). Amparado en la descomposición de noticias (*News*) analizada en la Sección 3.3.3.3, se puede saber el impacto de cada nuevo dato sobre la revisión final del PIB. Al filtrar el error de predicción contemporáneo de cada serie a través de la Ganancia de Kalman Suavizada, se dota al sistema de una total transparencia estructural. Es posible determinar cuantitativamente si la corrección en la perspectiva económica ha sido causada por una innovación negativa en las encuestas de opinión o por una expansión imprevista en los registros de empleo, un requisito fundamental para la toma de decisiones dentro de un entidad bancaria.

Parte II

APLICACIÓN PRÁCTICA

Capítulo 4

Panel de datos y pre-procesado

En este capítulo se aborda la aplicación práctica de la metodología introducida a lo largo del trabajo, con el objetivo de construir la base de un modelo econométrico que permita a la entidad obtener una estimación anticipada (*nowcast*) del Producto Interior Bruto (PIB) español. Para ello, se parte de un panel de datos consolidado compuesto por la serie del PIB y 31 covariables macroeconómicas, seleccionadas tras un estudio exhaustivo y validadas por el criterio de los expertos en análisis macroeconómico de la entidad.

Como se evidenciará mediante el análisis exploratorio, la dinámica histórica de la serie de interés, el PIB, suele dibujar una trayectoria suave que, sin embargo, se ve truncada por caídas abruptas durante perturbaciones severas, como la crisis financiera de 2008 o el colapso generado por la pandemia de la COVID-19 en 2020. Es precisamente ante estos cambios bruscos de nivel y puntos de giro donde un Modelo Factorial Dinámico justifica plenamente su uso: al incorporar un panel de covariables disponibles con mayor frecuencia (mensual) y menor retardo de publicación, el modelo extrae información adelantada sobre la continuidad y tracción del ciclo económico antes de la publicación oficial, ganando un tiempo vital para la planificación estratégica de la organización empresarial.

El análisis se estructura en dos fases secuenciales. En la primera, se describe el panel de variables, la serie del PIB de referencia y el protocolo de extracción automatizada de las fuentes oficiales. En la segunda, se detalla el procesamiento estadístico mediante el algoritmo X-13-ARIMA-SEATS para corregir la estacionalidad, así como las transformaciones necesarias para garantizar la estacionariedad débil y la estandarización final del panel.

4.1. Panel de datos

A continuación, se realizará un análisis exploratorio del panel de variables formado por la variable de interés y los indicadores económicos seleccionado junto con los criterios cualitativos que se han seguido para su introducción al set de variables. Además, en esta sección se explicará detalladamente el proceso de automatización de descarga de las series que se ha llevado a cabo en el presente proyecto.

4.1.1. Serie del PIB

La principal variable macroeconómica de referencia dentro del panel de datos es el índice del Producto Interior Bruto (PIB) a precios constantes (base 2020). Esta serie proviene de la Contabilidad Nacional Trimestral publicada por el Instituto Nacional de Estadística (INE) y se incorpora al estudio ya corregida de efectos estacionales y de calendario. Se trata de una serie de frecuencia trimestral, cuyo período de análisis abarca desde el primer trimestre de 1995 hasta el primer trimestre de 2026.

Dada la orientación de este trabajo hacia el análisis de coyuntura económica en tiempo real (*nowcasting*), es fundamental considerar el calendario de difusión oficial: a los 30 días de concluir el trimestre de referencia el INE publica la estimación del "Avance del PIB" (en adelante, primera estimación o

dato *flash*), mientras que a los 90 días se divulga el dato oficial completo de la Contabilidad Nacional, en adelante segunda estimación.

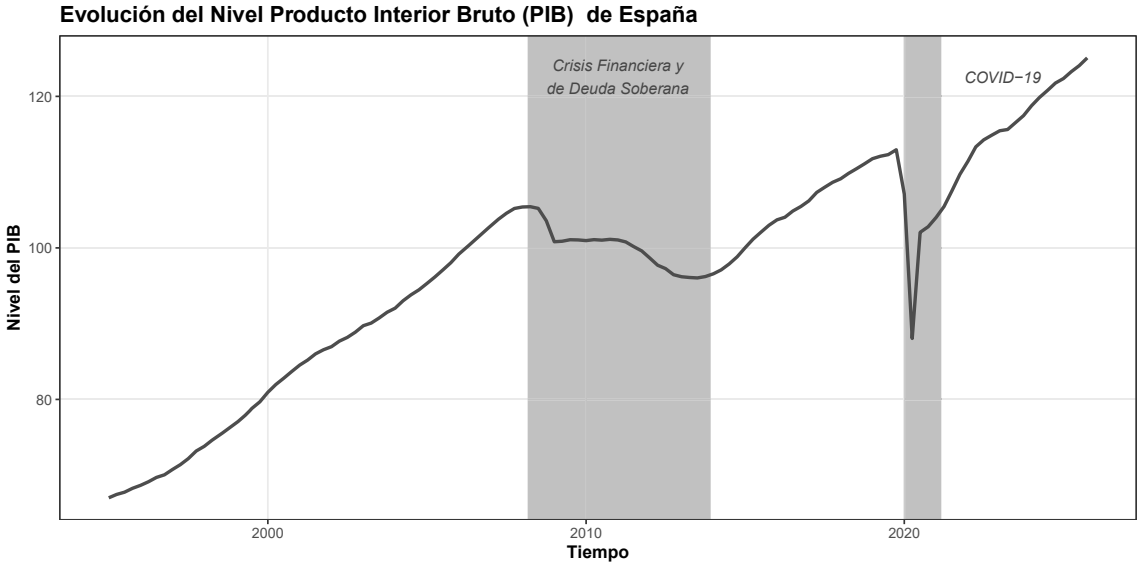


Figura 4.1: Evolución del PIB de España en nivel con los periodos de recesión sombreados (1995-2026)
.

La Figura 4.1 ilustra la evolución histórica de la serie en niveles. Las áreas sombreadas en gris delimitan las fases de recesión económica experimentadas por España en el horizonte temporal 1995-2026.[1] Como se observa en la figura, la serie refleja con fidelidad las contracciones de la actividad productiva durante los episodios de crisis macroeconómica.

El último dato disponible con el que se cierra la muestra corresponde al primer trimestre de 2026 (Q1-2026), el cual parece consolidar la trayectoria expansiva de los últimos años. Cabe destacar la suavidad relativa que mantiene la tendencia histórica en la mayor parte de la muestra, incluso durante la Gran Recesión de 2008-2013. No obstante, la crisis de la COVID-19 destaca como un choque exógeno sin precedentes dentro de la serie, generando un shock en forma de V. Este comportamiento anómalo responde a las severas medidas de confinamiento y a la paralización temporal de la actividad económica a escala global.

Con el fin de obtener una visión más completa de la evolución de esta variable, también se usa de apoyo el análisis de la serie histórica expresada en tasas de variación. Mientras que la representación en niveles es óptima para observar la tendencia de largo plazo, las transformaciones en diferencias permiten estabilizar la varianza, eliminar la no estacionariedad en media y evaluar el pulso coyuntural de la actividad en una unidad de medida homogénea expresada en porcentaje. Las trayectorias de la tasa intertrimestral, la tasa interanual y la variación anual acumulada se exponen detalladamente en la Figura 4.2:

La Figura 4.2 se define tal como:

- **Tasa de variación intertrimestral (porcentaje trimestral, % QoQ):** Cuantifica el crecimiento económico de un trimestre específico respecto al trimestre inmediatamente anterior. Formalmente, al trabajar con los datos ajustados de estacionalidad y calendario del INE, se

[1] Se adopta la definición convencional de recesión técnica (dos trimestres consecutivos de caída del PIB intertrimestral), en consonancia con la cronología oficial del Comité de Datación de Ciclos de la Asociación Española de Economía (AEE)(Asociación Española de Economía, 2015).

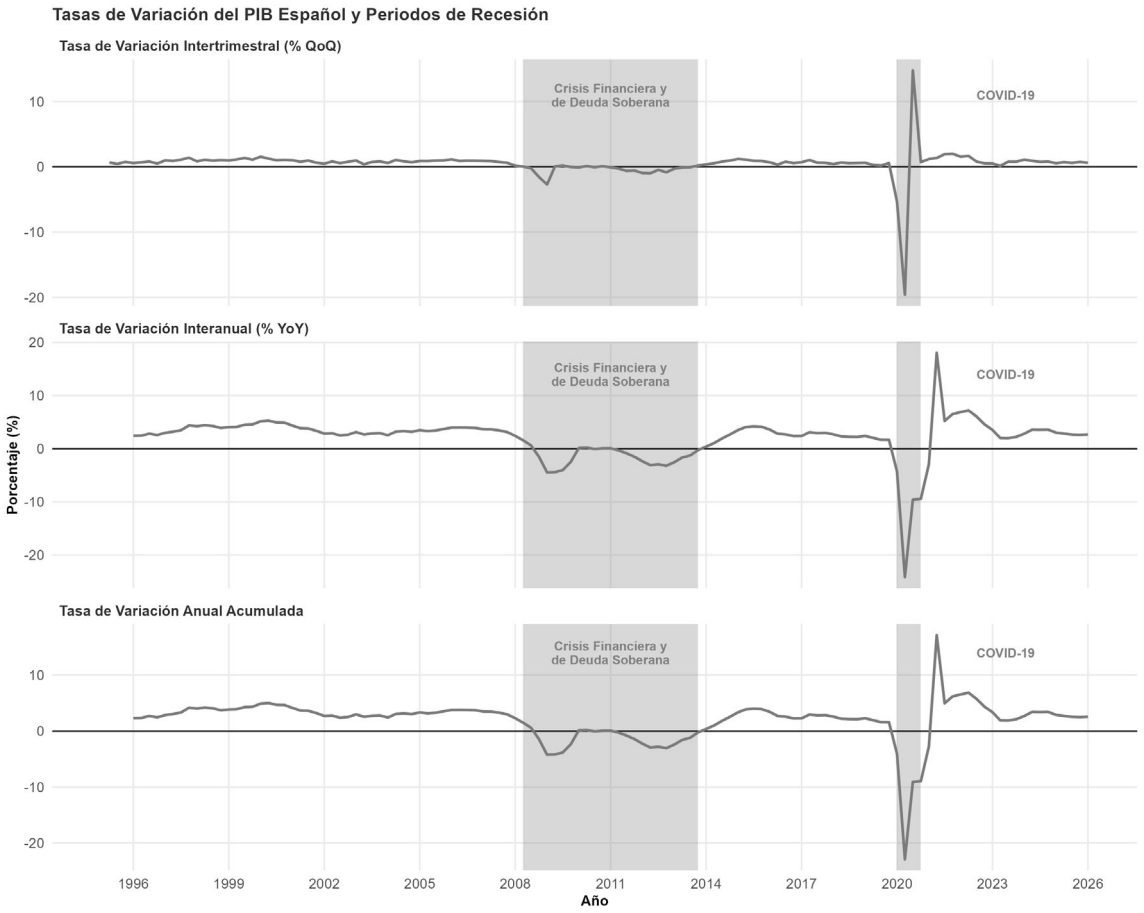


Figura 4.2: Evolución del PIB de España en Var. Intertrimestral, Var. Interanual y Var. Anual acumulada con los periodos de recesión sombreados (1995-2026).

.

calcula mediante la primera diferencia logarítmica regular de la serie:

$$\Delta y_t = \ln(\text{PIB}_t) - \ln(\text{PIB}_{t-1}) \approx \frac{\text{PIB}_t - \text{PIB}_{t-1}}{\text{PIB}_{t-1}}.$$

Esta transformación es la que garantiza la propiedad estacionariedad débil para el PIB, integrada de orden cero o $I(0)$, requerida para la entrada en el Modelo Factorial Dinámico. Su comportamiento en la figura se caracteriza por exhibir una alta volatilidad y capturar de forma inmediata los puntos de inflexión del ciclo, aislando el impulso de muy corto plazo de la actividad económica.

- **Tasa de variación interanual (porcentaje anual, % YoY):** Mide el crecimiento de un trimestre respecto al mismo trimestre del año anterior, aislando cualquier patrón estacional remanente. Su cálculo se formaliza mediante la diferencia logarítmica estacional de orden cuatro $(t-4)$:

$$\Delta_4 y_t = \ln(\text{PIB}_t) - \ln(\text{PIB}_{t-4}) \approx \frac{\text{PIB}_t - \text{PIB}_{t-4}}{\text{PIB}_{t-4}}.$$

En el análisis exploratorio, esta serie suaviza el ruido transitorio presente en la tasa intertrimestral, ofreciendo una perspectiva más robusta del pulso económico.

- **Tasa de variación anual acumulada:** Representa el crecimiento del año acumulado (promedio móvil de los últimos cuatro trimestres) en comparación con el mismo periodo del año precedente. Al promediar los flujos, opera un alisado que suaviza la serie, permitiendo evaluar la inercia macroeconómica de fondo de cada ejercicio.

El análisis comparativo de estas tres tasas a lo largo del horizonte 1995-2026 permite contrastar la profundidad y la morfología de las crisis (shocks) que ha sufrido la economía española de forma coherente con la evidencia gráfica:

Durante la crisis financiera de 2008-2013, la recesión se manifiesta como un fenómeno de destrucción persistente de doble suelo o doble caída (perfil en "W"). La tasa intertrimestral (primer panel) encadena variaciones negativas consecutivas que tocan fondo en el $-3{,}1\,\%$ a principios de 2009, para luego experimentar una breve meseta de estancamiento en torno al $0\,\%$ en 2010 y sufrir una segunda caída menos profunda (cercana al $-1{,}5\,\%$) en 2012.

Este patrón se traslada de forma amplificada a la tasa interanual (segundo panel), dibujando dos mínimos muy pronunciados: el primero en 2009, donde la tasa interanual se desploma hasta el $-4{,}5\,\%$, y el segundo en 2012, donde vuelve a hundirse hasta el $-3{,}5\,\%$. La tasa anual acumulada (tercer panel) replica esta misma estructura temporal, pero con transiciones mucho más suaves y un ligero retardo, reflejando la inercia del promedio móvil anual ante la persistencia de la recesión.

Por el contrario, la crisis de la COVID-19 en 2020 presenta una morfología radicalmente distinta, caracterizándose por ser un choque exógeno fulminante y de varianza extrema (perfil en "V"). La tasa intertrimestral registra una caída vertical sin precedentes en el segundo trimestre de 2020, desplomándose hasta el $-17{,}8\,\%$, seguida de un rebote histórico del $+16{,}8\,\%$. en el tercer trimestre del mismo año al retirarse las restricciones sanitarias estrictas.

Este colapso puntual se traduce en las tasas interanual y acumulada en un hundimiento drástico que rompe el suelo del $-20\,\%$, situándose en el $-21{,}5\,\%$ interanual en el segundo trimestre de 2020. Asimismo, el rebote técnico posterior genera un fuerte "efecto escalón" o base en 2021, impulsando la tasa interanual hasta un pico del $+17{,}7\,\%$ por la mera comparación estadística frente al trimestre colapsado del año anterior.

Finalmente, el extremo derecho de los tres paneles confirma que, tras absorber la volatilidad de la pandemia, la economía española estabiliza su ritmo de crecimiento hacia el primer trimestre de 2026. La convergencia de las tasas hacia niveles estables en torno al $2\,\%$ interanual constata que la transformación en diferencias logarítmicas ha purgado con éxito la raíz unitaria secular de los niveles, devolviendo un panel estacionario idóneo para la modelización.

4.1.2. Series de las Covariables

Para la selección de una batería inicial extensa de covariables se han tenido en cuenta un conjunto de criterios técnicos y macroeconómicos, apoyados tanto en la experiencia analítica como en referencias previas de instituciones oficiales; por ejemplo, el modelo Spain-STING del Banco de España (Arencibia Pareja et al., 2024; Camacho y Pérez-Quirós, 2011) y el modelo MIPred de la Autoridad Independiente de Responsabilidad Fiscal (AIReF) (AIReF, 2024; Cuevas y Quilis, 2015).

Los criterios de selección han sido los siguientes:

- **Representatividad sectorial**: Se han incluido variables que capturan la evolución de los principales pilares de la economía nacional (Consumo, Mercado Laboral, Industria, Servicios, Construcción, Sector Exterior y el entorno Financiero).
- **Longitud de la serie**: Se ha priorizado la inclusión de variables con una longitud de serie temporal extensa.
- **Frecuencia y granularidad**: Dado el objetivo de seguimiento continuo del PIB (variable trimestral), se ha exigido que las covariables presenten una frecuencia mínima mensual permitiendo actualizar la estimación a medida que fluye la información, con una única excepción, una variable trimestral en búsqueda de estabilidad en las estimaciones del modelo.
- **Retraso publicación**: Se han combinado variables de publicación casi inmediata o en tiempo real con aquellas de carácter más estructural que sufren mayores retrasos en su difusión oficial.
- **Naturaleza del dato**: El panel combina hard data (datos cuantitativos reales como afiliaciones a la Seguridad Social o consumo eléctrico) con soft data (datos cualitativos y de sentimiento económico).

A efectos metodológicos, el cierre de la base de datos completa se fijó el 21 de mayo de 2026, momento en el que se realizó la última extracción de las fuentes oficiales cuyo último dato disponible dentro de la muestra de las variables más adelantadas es de abril de 2026, el cual delimita el tamaño muestral del presente trabajo.

Este panel inicial nos deja un conjunto de variables representativas del conjunto de la economía española que ha sido corroborado por el experto en materia macroeconómica de la entidad. En la Tabla 4.1 se detalla este conjunto de información inicial, estructurado mediante agrupaciones sectoriales. Para facilitar su interpretación, la tabla detalla el identificador mnemotécnico asignado a cada variable (*Código*), su unidad de medida, su periodicidad -mensual (M) o trimestral (T)- y el retraso en su publicación oficial, expresado en días desde el cierre del periodo de referencia (*T+días*). Asimismo, se indica mediante una marca visual (✓) si la fuente original provee la serie con Corrección de Variaciones Estacionales y de Calendario (*CVEC*). Para clasificar el ámbito económico de los indicadores, se ha empleado la siguiente nomenclatura sectorial: Consumo (C), Construcción (Const), Industria (I), Mercado Laboral (L), Sector Exterior (E), Servicios (S), Sector Financiero (F) y entorno General (G). Por último, resulta pertinente destacar dos particularidades metodológicas en los datos de origen: la serie de consumo de energía eléctrica publicada por el MINECO ya incorpora una corrección del efecto temperatura y calendario, y las variables relativas al comercio exterior (exportaciones e importaciones) se presentan en volumen, al haber sido descontado previamente el Índice de Valor Unitario (IVU).

Tabla 4.1: Tabla resumen del Panel inicial de covariables seleccionadas para el MFD

.

Variable	**Código**	**Sector**	**Unidades**	**Año Inicio**	**Retraso**	**Periodicidad**	**CVEC**
Matriculaciones de Turismos	MAT_TUR	C	Núm. vehículos	1989	T+1	M	–
Índice Comercio Minorista	ICM	C	Índice Base 2021	2000	T+28	M	✓
Indicador Confianza Consumo	ICC	C	Saldo resp.	1986	T+1	M	✓
Consumo Gasolina+Diesel	GASOLINA+DIESEL	C	Miles Toneladas	1996	T+35	M	–
Consumo Gas Natural	GAS	C	GWh (Gigavatios Hora)	2004	T+35	M	–
Grandes empresas: ventas Totales	VGE_TOT	C	Índice Base 2014	1996	T+40	M	✓
Grandes empresas: salarios	VGE_SAL	C	Índice Base 2014	1996	T+40	M	✓
Afiliación a la S.S.	AFIL_SS	L	Miles pers. (Media mensual)	2001	T+2	M	–
Paro Registrado	PARO	L	Miles pers.	1996	T+2	M	✓
Ocupados	OCU	L	Miles pers.	1996	T+28	T	✓
Indicador Confianza Sector Construcción	ICC_CONS	Const	Saldo resp.	1989	T+1	M	✓
Indice Producción Industrial Construcción	IPI_CONS	Const	Índice Base 2021	2005	T+36	M	✓
Compraventa de vivienda	CP_VIV	Const	Núm. operac.	2007	T+50	M	–
Licitación Oficial	LICIT	Const	Millones euros	1989	T+55	M	–
Visados Obra Nueva	VIS	Const	Núm. visados	1991	T+60	M	–
Indice Producción Industrial	IPI	I	Índice Base 2021	1992	T+36	M	✓
Consumo Aparente de cemento	CEM	I	Miles Toneladas	1964	T+15	M	–
Consumo Energía Eléctrica	CONS_ELEC	I	GWh (Gigavatios hora)	1981	T+1	M	✓
Matriculaciones de Vehículos de Carga	MAT_CARGA	I	Núm. vehículos	1993	T+1	M	–
Índice PMI Manufacturas	PMI_MANU	I	Índ. Difusión	2000	T+3	M	✓
Exportaciones de Bienes	EXP_BIENES	E	Millones euros	1991	T+50	M	–
Importaciones de Bienes	IMP_BIENES	E	Millones euros	1991	T+50	M	–
Transporte Marítimo de Mercancías	TRAF_PORT	E	Toneladas	1997	T+55	M	–

Continúa en la siguiente página...

Tabla 4.1: Tabla resumen del Panel inicial de covariables (continuación)

.

Variable	Abrev.	Sector	Unidades	Año Inicio	Retraso	Periodicidad	CVEC
Gasto de Turistas Internacionales	GAS_TUR	E	Toneladas	2015	T+33	M	–
Indice cifra negocios Servicios	ICN_SERV	S	Índice Base 2021	2000	T+45	M	✓
Pernoctaciones Viajeros Totales	PERNOCT_TOT	S	Núm. pernoct.	1999	T+23	M	–
Índice PMI Servicios	PMI_SERV	S	Índ. Difusión (50 neutro)	2000	T+3	M	✓
Tráficos de pasajeros en aeropuertos	PASAJ_AER	S	Núm. pasajeros	1990	T+12	M	–
Entrada de viajeros: turistas internacionales	TURIS_INT	S	Núm. turistas	2015	T+33	M	–
Financiación Familias/Hogares	FIN_FAM_NUEVAS	F	Millones euros	2003	T+35	M	–
Indicador Sentimiento Económico	ISE	G	Índice	1987	T+3	M	✓

Continúa en la siguiente página...

Las series recopiladas son de carácter abierto y proceden de organismos públicos e instituciones oficiales. Las fuentes primarias de información empleadas en este estudio son las siguientes:

- INE (Instituto Nacional de Estadística).
- MINECO (Ministerio de Economía, Comercio y Empresa).
- MTES (Ministerio de Trabajo y Economía Social) y TGSS (Tesorería General de la Seguridad Social).
- AEAT (Agencia Estatal de Administración Tributaria) y Departamento de Aduanas.
- BdE (Banco de España).
- Comisión Europea (Eurostat).
- Otras instituciones públicas y operadoras: Red Eléctrica de España (REE), la Corporación de Reservas Estratégicas de Productos Petrolíferos (CORES) y el Ministerio de Transportes y Movilidad Sostenible (MTMS).

A continuación, se justificará de forma detallada la selección del panel inicial de covariables a través del crtierio de representatividad de los sectores económicos. Los razonamientos se fundamentan en los criterios presentados en el inicio de esta sección y el indispensable criterio experto que suministró el responsable de análisis macroeconómico de la entidad.

Asimismo, se detallarán las transformaciones aplicadas a ciertas variables para garantizar su consistencia con la medición del PIB a precios constantes y corregir distorsiones coyunturales, como las sufridas durante la pandemia en las afiliaciones a la seguridad social.

- **Consumo:** El consumo privado es el pilar fundamental de la economía española, representando históricamente en torno al 58-60 % del PIB desde la óptica de la demanda. Resulta evidente la estrecha relación entre los indicadores de gasto de los hogares y la evolución del ciclo económico general. Por esta razón, se ha incluido una amplia batería de variables que capturan distintas dimensiones del consumo. Se incorporan *hard data* como el Índice de Comercio Minorista (ICM), las Matriculaciones de Turismos y el Consumo de Combustibles, junto con indicadores de expectativas como el Indicador de Confianza del Consumidor (ICC). Adicionalmente, se incluyen las series de Ventas Totales y Salarios de las Grandes Empresas publicadas en la Agencia Tributaria. Cabe destacar una corrección sobre la variable de Retribución Bruta (VGE_SAL): dado que se publica en términos nominales, se ha procedido a deflactarla [2] utilizando el Índice de Precios al Consumo (IPC). Este paso es estrictamente necesario para limpiar el efecto de la inflación en la masa salarial y mostrar la evolución del poder adquisitivo real, haciéndola congruente con un PIB medido a precios constantes.

- **Mercado Laboral:** Los indicadores de este sector se caracterizan por su alta correlación procíclica y su nulo o escaso retraso en la publicación (el Paro y la Afiliación se publican a los dos días de finalizar el mes de referencia). Esta puntualidad es una característica importante de estas variables, ya que sumado a su importancia relativa dentro del comportamiento de la economía permite anticiparlo antes de la llegada de otras variables *hard data*. La inclusión de la Afiliación a la Seguridad Social (AFIL_SS) ha requerido de una corrección producto de la pandemia reciente. Entre marzo de 2020 y marzo de 2022, la irrupción de la COVID-19 obligó al despliegue masivo de los Expedientes de Regulación Temporal de Empleo (ERTE). Estadísticamente, los trabajadores en ERTE seguían contabilizando como afiliados, pero en la práctica no generaban valor añadido. Para solventar este sesgo, la serie original ha sido depurada restando el número de trabajadores en ERTE durante dicho periodo, obteniendo así una medida de empleo efectivo o real. Se incluye también la serie de Ocupados de la EPA (CVEC), la cual, a pesar de su frecuencia trimestral, se cree que por su importancia histórica merece ser añadida en el panel.

- **Industria:** Aunque el peso del sector industrial en el PIB de España, desde el punto de vista de la oferta (Valor Añadido Bruto, en adelante VAB) se sitúa en torno al 16 %, se trata del sector con mayor varianza cíclica de la economía. Su alta sensibilidad a los choques económicos lo convierte en un excelente indicador coincidente. Se incluyen variables determinantes como el Índice de Producción Industrial (IPI). . Una variable de especial relevancia para el seguimiento de la actividad es el Consumo de Energía Eléctrica (CONS_ELEC), dado que refleja la actividad de las fábricas en tiempo real. Para garantizar que esta serie actúe como un termómetro industrial puro, se utiliza la versión publicada por Red Eléctrica y el MINECO, la cual ya viene corregida de estacionalidad y depurada del efecto temperatura, aislando los picos de demanda debidos a olas de frío o calor. Se complementa el sector con el índice adelantado PMI [3] de Manufacturas que permite obtener una alerta temprana del funcionamiento de este sector.

- **Construcción:** El sector de la construcción y la promoción inmobiliaria, pese a no haber recuperado el peso relativo sobre el PIB previo a la crisis financiera de 2008, en torno a un 11 % en su punto más álgido, conserva un efecto multiplicador fundamental sobre el empleo y la inversión agregada. Su inclusión en el panel resulta motivada por este contexto y justificada y defendida por el experto en macroeconomía de la entidad que debe añadirse alguna variable bajo un criterio de representatividad sectorial e importancia histórica.

[2] Deflactar es un término económico que consiste en transformar una magnitud monetaria expresada a precios nominales (corrientes) en otra a precios reales, eliminando el efecto de la inflación

[3] El Índice de Gestores de Compras (PMI, por sus siglas en inglés) es un indicador económico adelantado elaborado a partir de encuestas mensuales a directivos de empresas. Refleja su percepción sobre la evolución de variables clave del negocio (como nuevos pedidos, producción o contratación). Su interpretación se basa en un umbral de 50 puntos: valores superiores indican una expansión de la actividad respecto al mes anterior, mientras que registros inferiores a 50 señalan una contracción.

No obstante, los analistas macroeconómicos coinciden sistemáticamente en señalar a la construcción como uno de los sectores más opacos y complejos de monitorizar en tiempo real. Existe una notable escasez de series estadísticas de alta frecuencia que cuenten con un retardo de publicación aceptable para el ejercicio del *nowcasting*. Esta deficiencia estructural se explica por la propia idiosincrasia de la actividad: los extensos plazos de ejecución de las obras y la pesada burocracia administrativa (certificaciones, aprobaciones municipales, firmas notariales) generan una fricción temporal que se traduce en un evidente subdesarrollo estadístico.

Los escasos datos disponibles suelen presentar dinámicas extremadamente volátiles y dependientes de decisiones discrecionales o institucionales que introducen ruido en la serie, como es el caso de la Licitación Oficial, fuertemente sujeta a los ritmos y presupuestos de la contratación pública. Para compensar estas carencias y dotar de equilibrio al sector dentro del modelo, la estrategia de selección ha consistido en combinar métricas de producción pura (como el Índice de Producción de la Construcción) con indicadores adelantados del inicio del ciclo constructivo y de demanda, como los Visados de Obra Nueva y la Compraventa de Viviendas. Finalmente, este bloque se apuntala con el componente cualitativo del Indicador de Confianza del Sector (ICC_CONS), mitigando así el severo decalaje temporal de los registros oficiales.

- **Sector Exterior:** Desde el punto de vista macroeconómico, las Exportaciones y las Importaciones son componentes directos de la ecuación contable del PIB. Es necesario incluirlas para capturar el efecto de la demanda externa y los cuellos de botella globales. Las series de importaciones y exportaciones de bienes publicadas por Aduanas se presentan originalmente en precios corrientes. Para introducirlas en el MFD sin distorsionar la señal del PIB real, estas series han sido deflactadas descontando el Índice de Valor Unitario (IVU). A través de esta transformación, se aísla el volumen físico de las transacciones internacionales frente a la volatilidad de los precios en las aduanas. El sector se complementa con métricas de turismo internacional (Gasto y Entradas), de extrema relevancia para el PIB español, y el tráfico de mercancías en puertos y aeropuertos.
- **Servicios:** El sector terciario representa el bloque más pesado de la economía española (en torno al 68 % del VAB). Su comportamiento sostenido ancla la tendencia del PIB general. Se han seleccionado los indicadores de facturación del sector (ICN_SERV) y la encuesta PMI de Servicios, esta última valorada por su gran capacidad de predicción adelantada al basarse en las expectativas de los gestores de compras. Por su relación directa con la "industria" turística nacional, se incorpora la pernoctación de viajeros totales.
- **Entorno General y Financiero:** Finalmente, para capturar el acceso al capital y el sentimiento transversal que no puede ser acotado a un solo sector, se incluyen variables sistémicas. Se incorpora el Indicador de Sentimiento Económico (ISE) sintetizado por la Comisión Europea, que actúa como un barómetro general de expectativas. Del mismo modo, se añade el volumen de Financiación a Familias (Nuevas operaciones), variable que permite adelantar la propensión al consumo duradero y la inversión residencial, y que captura el mecanismo de transmisión de la política monetaria.

4.1.3. Análisis Descriptivo de las variables

Una vez hecha una definición introductoria de las covariables que integran el panel, resulta de interés evaluar la relación existente con la variable objetivo. Para analizar de forma preliminar el comovimiento entre series se aborda mediante un análisis de correlaciones cruzadas, evaluando una malla de rezagos y adelantos temporales de hasta tres trimestres ($\tau \in \{-3, \dots, +3\}$) frente al PIB.

Formalmente, el coeficiente de correlación cruzada muestral entre una covariable seleccionada x_t y el PIB real y_t para un desfase temporal τ se calcula mediante la covarianza cruzada estandarizada:

$$\rho_{xy}(\tau) = \frac{\gamma_{xy}(\tau)}{\sqrt{\gamma_{xx}(0)\gamma_{yy}(0)}},$$

donde $\gamma_{xx}(0)$ y $\gamma_{yy}(0)$ representan las varianzas muestrales de cada serie respectivamente, mientras que la covarianza cruzada muestral $\gamma_{xy}(\tau)$ para un tamaño muestral T se define operativamente como:

$$\gamma_{xy}(\tau) = \begin{cases} \frac{1}{T}\sum_{t=1}^{T-\tau}(x_{t+\tau} - \bar{x})(y_t - \bar{y}) & \text{si } \tau \geq 0 \\ \frac{1}{T}\sum_{t=1-\tau}^{T}(x_{t+\tau} - \bar{x})(y_t - \bar{y}) & \text{si } \tau < 0, \end{cases},$$

siendo $\bar{x}$ y $\bar{y}$ las medias muestrales de las series correspondientes en el horizonte temporal analizado.

Bajo esta formulación, el signo y el orden de magnitud del desfase temporal τ determinan la dirección y el carácter cíclico de la relación de la siguiente manera:

- **Indicador adelantado ($\tau < 0$):** Un coeficiente elevado y estadísticamente significativo para valores negativos de τ implica que las realizaciones pasadas de la covariable ($x_{t-\tau}$) covarían con el PIB contemporáneo (y_t), lo que significa que el indicador tiene una capacidad de anticipación.
- **Indicador coincidente ($\tau = 0$):** Representa la relación contemporánea síncrona entre ambas variables dentro de un mismo trimestre.
- **Indicador retardado ($\tau > 0$):** Refleja que la dinámica PIB precede a los movimientos de la covariable analizada, comportándose esta última como un rezago del ciclo económico.

Los coeficientes estimados para el conjunto de indicadores seleccionados se exponen en el mapa térmico de la Figura 4.3:

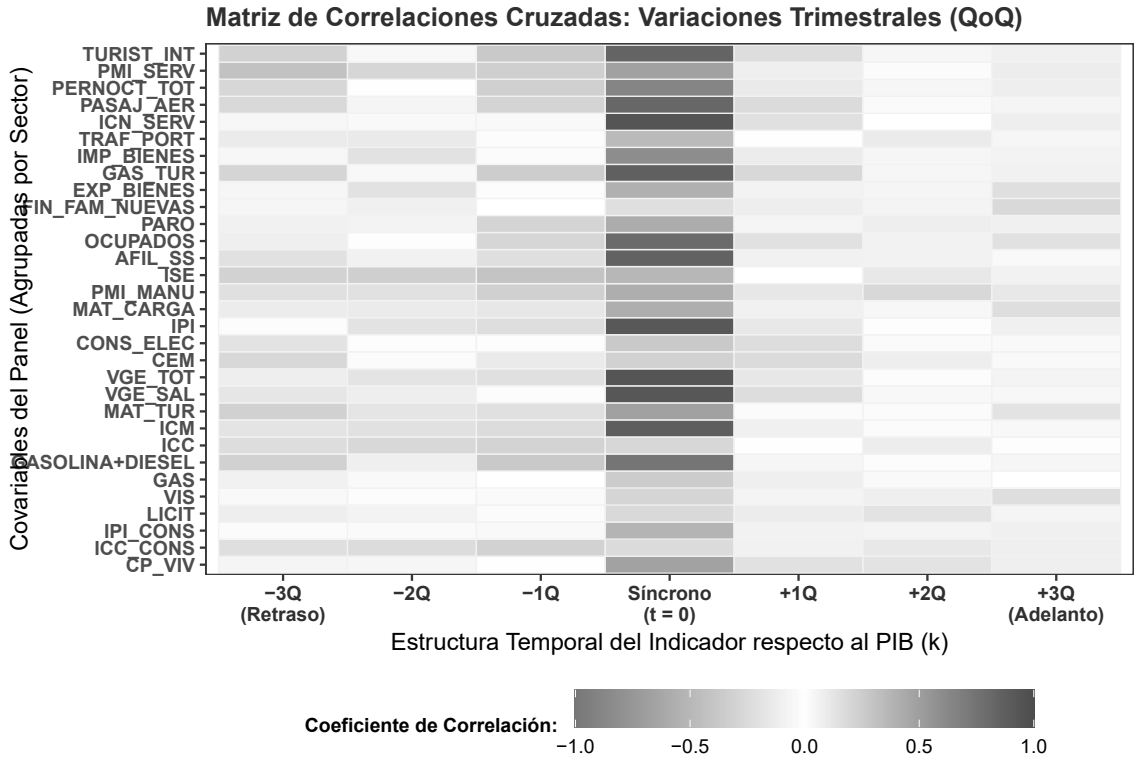


Figura 4.3: Mapa de calor de correlaciones cruzadas de los indicadores frente al PIB. Elaboración propia.

.

Este análisis descriptivo de las variables a través de las correlaciones del panel permite extraer algunas conclusiones al respecto de la relación de las covariables frente a la variable de interés:

1. **Relación contemporánea entre las variables:** El mapa de calor evidencia que la máxima densidad de comovimiento se concentra de forma global en la columna central de sincronización contemporánea ($\tau = 0$). En el enfoque intertrimestral (QoQ), los principales motores cuantitativos de la economía real española exhiben una vinculación contemporánea muy alta: la Cifra de Negocios del Sector Servicios (`ICN_SERV`) registra un coeficiente de 0,941, las Ventas Totales de Grandes Empresas (`VGE_TOT`) un 0,937, y el Índice de Producción Industrial (`IPI`) un 0,917. Al evaluar los desplazamientos hacia retardos o adelantos ($\tau \neq 0$), estas magnitudes experimentan una reducción. Esta distribución de valores constituye la justificación para que en la ecuación de medida del MFD se asuma una relación estrictamente contemporánea entre los indicadores y el factor común latente.

2. **Relación con las variables de Soft Data:** En contraposición, las variables cualitativas procedentes de encuestas de opinión muestran una notable capacidad para capturar de forma temprana el sentimiento de la economía. Al analizar las variaciones trimestrales (QoQ), los índices de difusión y expectativas, como el `PMI_MANU` (con una correlación síncrona de 0,413) y el `PMI_SERV` (0,485), junto con el Indicador de Confianza del Consumidor (`ICC`), muestran una presencia regular y estable a lo largo de las franjas temporales analizadas. Esta propiedad demuestra que las expectativas de los agentes económicos proporcionan al modelo una señal limpia y anticipada del pulso de la actividad antes de que se consoliden y publiquen los registros contables oficiales.

3. **Comportamiento contracíclico:** La variable de Paro Registrado (`PARO`) se ratifica como el único indicador contracíclico del sistema, arrojando un coeficiente contemporáneo de $-0{,}587$ (QoQ). Por su parte, los indicadores asociados a la construcción reflejan los extensos plazos de maduración de la inversión inmobiliaria y la obra pública. En el enfoque intertrimestral, variables como la Compraventa de Viviendas (`CP_VIV`) marcan una correlación contemporánea significativa de 0,480, mientras que el Índice de Producción de la Construcción (`IPI_CONS`) se sitúa en 0,380. Estos desfases y dinámicas específicas confirman que el sector constructivo responde a ciclos de maduración más largos, donde los movimientos del presente condicionan de forma de fondo la actividad económica de los trimestres posteriores.

4.2. Automatización de descarga de series del modelo

Para asegurar el correcto seguimiento de la economía española a través del *nowcasting*, se exige que el flujo de actualización de las variables hacia el Modelo Factorial Dinámico se ejecute sin intervención manual, evitando errores e inconsistencias y en el menor tiempo posible. Con el fin de dotar al modelo de un carácter plenamente operativo y reproducible, se ha programado un *script* en lenguaje R que automatiza la ingesta, limpieza, agregación y exportación de todo el panel de datos a un archivo Excel.

Dado que los distintos organismos públicos no comparten un estándar único de difusión, la arquitectura de este proceso de extracción se ha diseñado empleando diversas técnicas informáticas adaptadas a la naturaleza de cada fuente, las cuales se detallan a continuación.

Para la extracción de variables provenientes del Instituto Nacional de Estadística (INE), como el Índice de Producción Industrial o las pernoctaciones turísticas, se ha explotado su interfaz de programación de aplicaciones (API). El INE provee el servicio *API JSON TEMPUS*, que permite acceder a las series temporales dinámicamente mediante la consulta de URLs. La estructura general de la petición es:

```
https://servicios.ine.es/wstempus/js/ES/DATOS_TABLA/{id_tabla}?nult={n_datos}
```

En esta estructura, el parámetro `{id_tabla}` hace referencia al código numérico identificativo único de la serie en la base de datos del INE. Por su parte, el parámetro `nult` especifica la cantidad de últimos períodos de datos históricos que se desean recuperar. De forma alternativa, la API permite modificar el parámetro `date` por si se prefiere acotar la consulta a un rango de fechas específico.

El formato de salida nativo es JSON (*JavaScript Object Notation*). Para su tratamiento, se utiliza la librería `jsonlite` (Ooms, 2014), que decodifica la respuesta web y la transforma en listas anidadas

dentro de R. Posteriormente, se extraen los vectores de valores y fechas, y se convierten en objetos de series temporales (`ts`). Desde esta fuente se obtienen directamente variables fundamentales como el Índice de Cifra de Negocios en la Industria (`61709`), el Índice de Producción Industrial bruto (`60280`), el Gasto de turistas internacionales (`10828`), la Entrada de viajeros (`10821`) y las pernoctaciones hoteleras.

A modo ilustrativo, el siguiente bloque de código muestra cómo itera sobre las series del INE, extrae el vector de valores y lo convierte automáticamente en un objeto de serie temporal:

```
series_ine <- c("10828", "10821", "61709", "60280")
url_ine <- paste0("https://servicios.ine.es/wstempus/js/ES/DATOS_TABLA/",
series_ine, "?nult=10000000")

datos_ine_ts0 <- c()
for (i in 1:length(series_ine)){
  json <- jsonlite::fromJSON(url_ine[i])
  # Extraccion de valores y conversion a serie temporal (ts)
  serie <- na.omit(ts(as.numeric(json$Data[[1]][dim(json$Data[[1]])[1]:1,]$Valor),
  start = c(json$Data[[1]]$Anyo[dim(json$Data[[1]])[1]],
  json$Data[[1]]$FK_Periodo[dim(json$Data[[1]])[1]]),
  freq = max(json$Data[[1]]$FK_Periodo)))
  datos_ine_ts0 <- cbind(datos_ine_ts0, serie)
}
```

Para las instituciones que difunden sus bases de datos mediante archivos masivos comprimidos, como el Ministerio de Economía, Comercio y Empresa (MINECO) y la Agencia Estatal de Administración Tributaria (AEAT), se ha diseñado un protocolo centralizado de descarga y unificación analítica.

En el caso del **MINECO (a través de la Base de Datos Síntesis - BDSICE)**, el sistema realiza la descarga e indexación de los ficheros históricos de interés. Dado que la información sectorial se publica de forma fragmentada, el algoritmo realiza una consolidación relacional indexando todas las series bajo una jerarquía temporal común.

Por su parte, la rutina programada para la **AEAT** accede a los registros consolidados de la estadística de Grandes Empresas. El proceso automatiza la descompresión y el filtrado selectivo de los ficheros de texto plano para aislar e incorporar al panel general las series de Ventas Totales y de Retribuciones Salariales.

El siguiente ejemplo muestra la técnica empleada para la base de datos de la AEAT, requiriendo el uso de la librería `archive` para lidiar con el formato `.7z` y la lectura mediante la función `read.delim2`:

```
url_aeat <- "https://sede.agenciatributaria.gob.es/.../BdDVesge.7z"
destfile <- paste0(tempdir(),"//aeat.7z")
download.file(url_aeat, destfile, mode = "wb")

# Descompresion y lectura del fichero plano
archive::archive_extract(destfile, dir = paste0(tempdir(),"//aeat"))
datos <- read.delim2(paste0(tempdir(),"//aeat//Vesge_pub_indexada.txt"), sep=";")

# Filtrado de variables mediante codigos internos de la AEAT
ventas_ge <- ts(datos[which(datos$VARIABLE==2 & datos$LITERAL==29 &
datos$FREC==1 & datos$DATO1==1 &
datos$DATO2==2),]$VALOR,
start=c(1996,1), freq=12)
```

La mayor complejidad en la automatización se presentó en el acceso a los datos de la Corporación de Reservas Estratégicas de Productos Petrolíferos (CORES). A pesar de que las series de consumo de gas y combustibles residen en un hipervínculo estático con extensión `.xlsx`, el servidor web de la institución restringe las descargas directas mediante peticiones HTTP convencionales, exigiendo la validación de scripts en el lado del cliente y la gestión de sesiones activas.

Para sortear esta barrera, se implementó un sistema de *scraping* avanzado utilizando la librería `chromote`. Esta herramienta abre una sesión de Google Chrome y permite ejecutar funciones de JavaScript sobre la propia web. Así, el script carga la dirección seleccionada, localiza el nodo del botón de descarga a través de su selector CSS y automatiza su pulsación de forma programática:

```
library(chromote)
b <- ChromoteSession$new()
b$Browser$setDownloadBehavior(behavior = "allow", downloadPath = tempdir())
b$Page$navigate("https://www.cores.es/.../consumos-pp.xlsx")
Sys.sleep(2)

# Inyeccion de JavaScript para forzar la descarga
resultado <- b$Runtime$evaluate('
var boton = document.querySelector("input.btn");
if(boton) { boton.click(); }
')
Sys.sleep(2)
b$close()
```

Una vez el archivo `.xlsx` se descarga en el directorio temporal, R identifica el fichero modificado más recientemente, extrae las pestañas requeridas y desecha los totales agregados mediante rutinas de limpieza de cadenas de texto.

Para los datos provenientes de la oficina estadística europea, se aprovechó la existencia de la librería oficial `eurostat` en R (Lahti, Huovari, Kainu, y Biecek, 2017). Mediante la función `get_eurostat()`, se consultan directamente los *datasets* (como `ei_bssi_m_r2` para los indicadores de sentimiento), aplicando filtros para aislar la información exclusiva de España ("ES"), con ajustes estacionales ("SA") o estacionales y de calendario ("SCA"), unidad de medida (ej: Indice base 2021 ("I21"), además de la fecha que se necesite seleccionar. Un ejemplo de uso para obtener el IPI de la Construcción sería el siguiente:

```
library(eurostat)
prod_industrial <- get_eurostat("sts_copr_m", time_format = "date")
ipi_construccion <- prod_industrial %>%
filter(
geo == "ES",        # Espana
indic_bt == "PRD", # IPI
nace_r2 == "F",     # Construccion
s_adj == "SCA",     # Ajustado estacional y de calendario (CVEC)
unit == "I21"       # Indice base 2021 (o la mas actual disponible)
)
```

Por último, para el Banco de España y el Ministerio de Transportes, la información reside en archivos Excel estáticos con URLs fijas. Se utilizó la función base `download.file()` seguida de la lectura con `read_excel()` del paquete `readxl`. El reto analítico en esta fase consistió en estandarizar formatos de fecha heterogéneos (ej. "Ene 2024") utilizando expresiones regulares (`strsplit`) y lidiar con la presencia de caracteres atípicos (como guiones para referenciar datos ausentes) antes de forzar su tipología numérica.

Para que este *script* actúe de forma autónoma, la ejecución se ha de progamar mediante el paquete `taskscheduleR` (Wijffels, 2023). Esta herramienta interactúa directamente con el Programador de Tareas del sistema operativo Windows, permitiendo fijar rutinas de ejecución periódicas (diarias o semanales) coincidiendo con los días habituales de publicación de los datos.

Nota metodológica: Los datos propietarios de S&P Global (índices PMI manufacturero y de servicios) se integran en el modelo mediante la lectura directa desde un repositorio interno de la red de la entidad, dado que su acceso automatizado vía API se encuentra comercialmente restringido.

4.3. Pre-procesamiento de las series

Una vez que se tiene el panel de covariables presentado y justificada su selección inicial, se procede con el tratamiento de las series en aras de obtener un conjunto de series corregidas de estacionalidad, efecto calendario, sin tendencia y estandarizadas que permitan su entrada en el MFD de la forma adecuada. Con el fin de de resumir el procedimiento de preprocesado que se lleva a cabo para la entrada de datos en el MFD, se indica a través de la Figura 4.4 el algoritmo que se ha seguido para

ello. Además, se podrá observar a través de un ejemplo el procedimiento completo llevado a cabo sobre la variable Consumo de Energía Eléctrica.

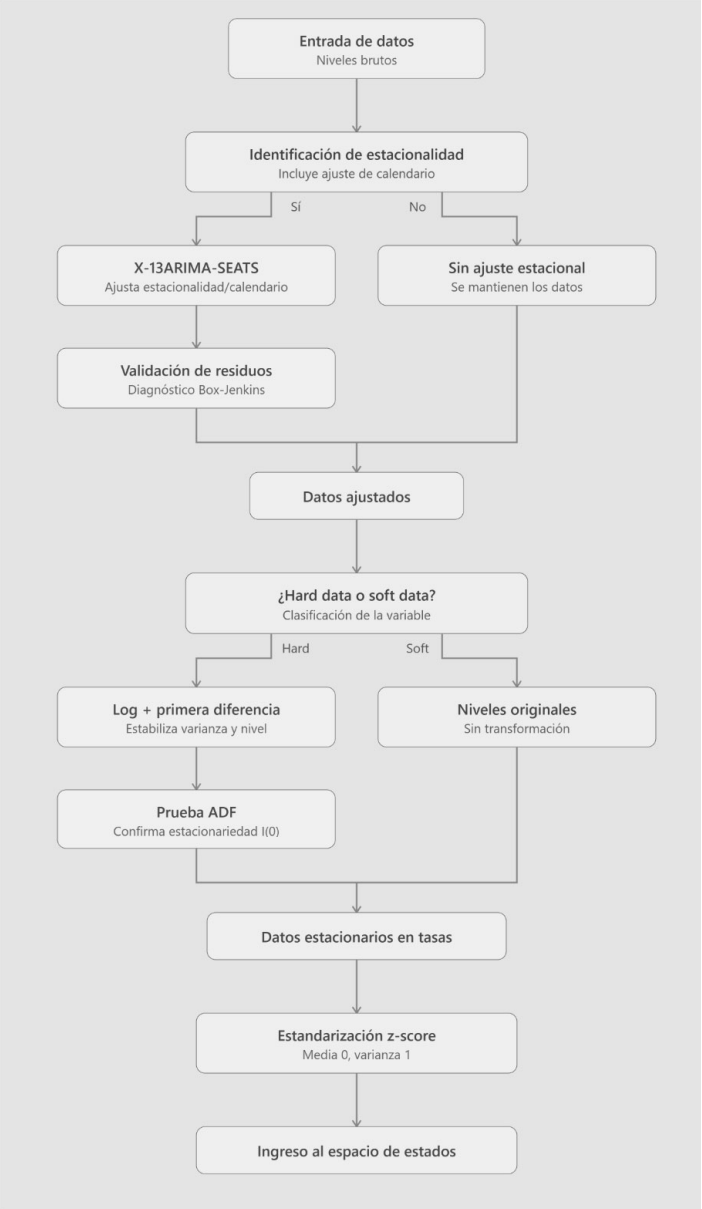


Figura 4.4: Algoritmo de preprocesado de series de tiempo para el MFD.Elaboración propia.

.

4.3.1. Detección de la Componente Estacional S_t

Detectar y aislar la componente estacional (S_t) (véase la Sección 2.2) dentro de las series temporales es un paso en la fase de preprocesamiento de los datos. Si no se detecta adecuadamente, las fluctuaciones intra-anuales (como las campañas de contratación de verano o el pico de consumo navideño) pueden contaminar la señal latente.

Si bien las fuentes oficiales ya proveen información sobre qué series han sido sometidas a una Corrección de Variaciones Estacionales y de Calendario (CVEC), como se observa en la Tabla 4.1, se ha realizado una comprobación a todas las variables por igual. Para ello, se ha empleado la librería `seastests` (Ollech, 2021), la cual somete a cada variable a una batería de contrastes de significación estadística para evaluar la presencia de raíces estacionales.

Con el objetivo de evitar sesgos derivados de la distribución subyacente de cada serie, la metodología evalúa cuatro contrastes estadísticos de distinta naturaleza, cuyos fundamentos matemáticos son los siguientes:

- **Test de Kruskal-Wallis (KW):** Es un contraste no paramétrico basado en rangos que evalúa si las muestras provienen de la misma distribución (Kruskal y Wallis, 1952). En este contexto, agrupa las observaciones por meses ($s = 12$ grupos) y verifica si la suma de los rangos de un mes difiere significativamente del resto.

- **Test de Friedman:** Actúa como una alternativa no paramétrica al ANOVA de medidas repetidas (Friedman, 1937). Organiza los datos en un diseño de bloques (donde los años son los bloques y los meses los tratamientos) y evalúa si existen diferencias sistemáticas en el nivel de la serie entre los distintos meses, aislando el efecto de la tendencia interanual.

- **Test QS (Qu-Hwang):** Es una variante del contraste de Ljung-Box diseñada específicamente para la estacionalidad (Lytras, Feldpausch, y Bell, 2007). Evalúa la presencia de autocorrelación significativa en los retardos estacionales de la serie (lags 12, 24, etc.). Un exceso de autocorrelación en dichos rezagos es un fuerte indicio de patrón estacional.

- **Test de Welch:** Es una adaptación del clásico Análisis de Varianza (ANOVA) para comparar las medias de los 12 meses, con la ventaja de ser robusto ante la heterocedasticidad (no asume que la varianza sea homocedástica en todos los meses), un rasgo muy común en variables macroeconómicas (Welch, 1951).

Dada la posibilidad de que los contrastes arrojen resultados contradictorios según naturaleza estocástica de la serie, el paquete proporciona una decisión unificada (*Combined*). Esta métrica se basa en un algoritmo de aprendizaje automático (*Random Forest*) entrenado por Ollech (2021) con decenas de miles de series reales y simuladas. El algoritmo evalúa de forma conjunta los p-valores de los cuatro test para emitir un veredicto robusto sobre si la serie requiere o no tratamiento, por ello aunque en alguna variable se de el caso de que 1 de los 4 contrastes indiquen que si existe componente estacional, el clasificador unificado puede identificar que se trata de un falso positivo e indicar de forma categórica que no es estadísticamente significativo.

En la Tabla 4.2 se exponen los resultados de esta batería de contrastes para el panel de covariables del modelo. La Columna Decisión (Combined) tiene 2 posibles respuestas, SI y NO, la primera indica que los test han detectado una componente estacional estadísticamente signficativa y la segunda que o bien no la ha detectado o no es estadísticamente significativa.

Los resultados de la tabla validan la categorización original de las fuentes. Las variables de "flujo" en bruto no tratadas previamente por los organismos (como Matriculaciones, Turismo, Exportaciones o Consumo de Energía) arrojan p-valores nulos, confirmando una estacionalidad determinista clara. En contraste, las variables procedentes de encuestas (*soft data* como el ISE, los PMIs y los Índices de Confianza) muestran p-valores cercanos a 1, algo que parecería lógico al tratarse de expectativas económicas. Del mismo modo, el algoritmo ratifica la limpieza aplicada por el INE y la AEAT en aquellas series catalogadas como CVEC (como el IPI, el Comercio Minorista o el PIB). Resulta sorprendente el caso de la variable Paro Registrado (`PARO`), considerada originalmente por la fuente CVEC, donde el contraste combinado muestra una conclusión unánime frente a los test individuales que parecen contradecirse. Mientras que los contrastes transversales (Kruskal-Wallis, Friedman y Welch) concluyen que no existen diferencias estacionales significativas ($p \approx 0{,}999$), el test de autocorrelación QS arroja un falso positivo ($p = 0{,}000$). Esta anomalía estadística ocurre debido a que el test QS confunde la

Tabla 4.2: Resultados del contraste múltiple de estacionalidad sobre el panel de covariables
.

		P-valores de los contrastes individuales			
Variable	**Decisión (Combined)**	**Kruskal-Wallis**	**Friedman**	**QS**	**Welch**
PARO	NO	0.998	0.699	0.000	0.939
IPI	NO	0.999	0.996	1.000	0.982
CEM	SÍ	0.000	0.000	0.000	0.000
CONS_ELEC	SÍ	0.000	0.000	0.000	0.000
ICM	NO	0.983	0.798	1.000	0.791
AFIL_SS	SÍ	0.000	0.000	0.003	0.000
MAT_CARGA	SÍ	0.000	0.000	0.000	0.000
VIS	SÍ	0.000	0.000	0.000	0.000
LICIT	SÍ	0.000	0.000	0.000	0.000
ICN_SERV	NO	0.998	0.976	1.000	0.911
MAT_TUR	SÍ	0.000	0.000	0.000	0.000
EXP_BIENES	SÍ	0.000	0.000	0.000	0.000
IMP_BIENES	SÍ	0.000	0.000	0.000	0.000
CP_VIV	SÍ	0.000	0.000	0.000	0.000
PASAJ_AER	SÍ	0.000	0.000	0.000	0.000
GAS_TUR	SÍ	0.000	0.000	0.000	0.000
TURIST_INT	SÍ	0.000	0.000	0.000	0.000
PERNOCT_TOT	SÍ	0.000	0.000	0.000	0.000
TRAF_PORT	SÍ	0.000	0.000	0.000	0.000
ICC_CONS	NO	0.146	0.144	1.000	0.104
ICC	NO	0.900	0.893	1.000	0.960
ISE	NO	0.955	0.940	1.000	0.890
IPI_CONS	NO	0.969	0.976	0.887	0.727
GASOLINA+DIESEL	SÍ	0.000	0.000	0.000	0.000
GAS	SÍ	0.000	0.000	0.000	0.000
FIN_FAM_NUEVAS	SÍ	0.000	0.000	0.000	0.000
VGE_TOT	NO	0.993	0.998	1.000	0.932
VGE_SAL	NO	0.962	0.892	0.400	0.932
PMI_SERV	NO	0.717	0.926	1.000	0.401
PMI_MANU	NO	0.372	0.329	1.000	0.327
OCU	NO	0.958	0.946	0.146	0.835
PIB	NO	0.944	0.631	1.000	0.430

fuerte inercia cíclica del mercado laboral español, como se comentó en la justificación sectorial, con un patrón estacional. El algoritmo unificado detecta que esta señal es espuria y descarta correctamente su tratamiento.

4.3.2. Corrección de estacionalidad y efecto calendario

A continuación, se prosigue con la corrección de estacionalidad y efectos calendario de las series del panel de covariables que han indicado a través de los contrastes realizados la necesidad de corrección de estacionalidad y efecto calendario. Esta corrección se llevará a cabo con el algoritmo presentado en la Sección 2.2.

El paquete `seasonal` desarrollado por Christoph Sax (Sax y Eddelbuettel, 2018) actúa través de su función principal `seas()`, que permite la automatización del proceso de modelado regARIMA, la selección óptima del esquema (aditivo o multiplicativo), la selección del motor de correción (X11 o SEATS) y la posibilidad de extracción de las componentes. Esta capacidad de automatización es fundamental en este trabajo debido a la necesidad de procesar iterativamente el amplio panel de indicadores que conformarán el modelo.

Como se desprende de la Tabla 4.3, las distintas pruebas para la corrección de las series han dado lugar a una parametrización heterogénea.

Un primer aspecto destacable es la estructura de los modelos *regARIMA* subyacentes. Denotados bajo la nomenclatura clásica $(p, d, q)(P, D, Q)$ (véanse los Apéndices A y A.25 para mayor profundidad analítica), los resultados revelan que la práctica totalidad del panel requiere tanto de una diferencia regular ($d = 1$) como de una diferencia estacional ($D = 1$). Más allá de la dinámica puramente estocástica, el módulo de regresión ha capturado el impacto de variables exógenas mediante los efectos de calendario. De forma coherente con la teoría económica, las series más expuestas a los patrones de consumo y la logística (`MAT_TUR`, `TRAF_PORT`) incorporan correcciones significativas por el efecto del día de la semana (*Weekday*) y la festividad de Semana Santa (*Easter*). Del mismo modo, en las series de volumen agregado, el algoritmo controla de forma automática el efecto bisiesto (*Leap Year*), evitando que un día adicional en febrero distorsione artificialmente las tasas de variación interanual. Desde el punto de vista metodológico, la categorización del panel en distintos niveles de intervención ha resultado crucial para preservar la integridad matemática del ajuste. Por un lado, la serie de licitación pública (`LICIT`) requirió el uso del filtro X-11 para acomodar su extrema volatilidad administrativa y evitar singularidades matriciales. Por otro lado, el tratamiento del clúster turístico (`PERNOCT_TOT`, `TURIST_INT`, `GAS_TUR`) supuso el mayor desafío algorítmico debido al colapso del sector en 2020, con algunos meses con 0 unidades. Frente a esta anomalía, la aplicación de un logaritmo desplazado ($y_t + 1$) sobre el motor X-11 ha permitido estabilizar la varianza heterocedástica y absorber matemáticamente los meses del confinamiento.

Como se indicó en la introducción de la sección, a modo de ejemplo se presenta esta fase del procesado de series a través de la variable Consumo de Energía Eléctrica (`CONS_ELEC`). Esta variable constituye un ejemplo pertinente dado que presenta una componente estacional muy acentuada –íntimamente ligada a los picos térmicos de demanda energética en los meses de invierno y verano– que resulta fácilmente identificable mediante un análisis exploratorio visual. Siguiendo la metodología descrita en la sección, esta variable ha sido procesada bajo la configuración del modelo "Paramétrico Regular". El motor paramétrico SEATS ha parametrizado la serie óptimamente bajo un esquema ARIMA $(1, 1, 1)(0, 1, 1)$. Se recuerda que esta variable ya viene corregida de efectos de temperatura y calendario, como se indicó en la Tabla 4.1.

Para observar de forma gráfica el antes y después de la serie tras el proceso de corrección de la componente estacional se presenta la Figura 4.5:

En primer lugar, la Figura 4.5 contrasta la serie bruta (línea negra) frente a la serie Corregida de Variaciones Estacionales y de Calendario o CVEC (línea roja). Se evidencia cómo el filtro de extracción de señales de SEATS ha suprimido el patrón repetitivo intra-anual (reduciendo la varianza estacional), devolviendo una señal estabilizada. Esta nueva serie limpia aísla la evolución pura de la tendencia-ciclo y el componente irregular, respetando los cambios estructurales de nivel y los puntos de inflexión de

Tabla 4.3: Especificaciones de los modelos de ajuste estacional (X-13ARIMA-SEATS)

.

Variable	**Método**	**Motor**	**ARIMA**	**Efectos Calendario**
CEM	Paramétrico Regular	SEATS	(2,1,1)(0,1,1)	Día de la semana, Semana Santa[1]
CONS_ELEC	Paramétrico Regular	SEATS	(1,1,1)(0,1,1)	Ninguno
AFIL_SS	Paramétrico Regular	SEATS	(2,1,0)(0,1,1)	Año bisiesto
MAT_CARGA	Paramétrico Regular	SEATS	(0,1,1)(0,1,1)	Año bisiesto, Día de la semana, Semana Santa[1]
MAT_TUR	Paramétrico Regular	SEATS	(1,1,1)(0,1,1)	Año bisiesto, Día de la semana, Semana Santa[1]
EXP_BIENES	Paramétrico Regular	SEATS	(1,1,1)(0,1,1)	Día de la semana, Semana Santa[1]
IMP_BIENES	Paramétrico Regular	SEATS	(3,1,1)(0,1,1)	Día de la semana, Semana Santa[1]
CP_VIV	Paramétrico Regular	SEATS	(0,1,0)(0,1,1)	Semana Santa[8]
PASAJ_AER	Paramétrico Regular	SEATS	(3,0,0)(0,1,1)	Año bisiesto, Día de la semana, Semana Santa[15]
TRAF_PORT	Paramétrico Regular	SEATS	(0,1,1)(0,1,1)	Día de la semana
GASOLINA+DIESEL	Paramétrico Regular	SEATS	(1,1,2)(0,1,1)	Año bisiesto, Semana Santa[1]
GAS	Paramétrico Regular	SEATS	(1,1,1)(0,1,1)	Día de la semana
VIS	Paramétrico Regular	SEATS	(2,1,2)(0,1,1)	Día de la semana, Semana Santa[8]
FIN_FAM_NUEVAS	Paramétrico Regular	SEATS	(1,1,1)(1,1,1)	Día de la semana, Semana Santa[1]
LICIT	Intervención Exógena por Volatilidad	X-11	(0,1,1)(0,1,1)	Día de la semana
GAS_TUR	Tratamiento Aditivo por Shock	X-11	(1,1,0)(0,1,1)	Año bisiesto, Día de la semana, Semana Santa[1]
TURIST_INT	Tratamiento Aditivo por Shock	X-11	(1,1,0)(1,1,1)	Año bisiesto, Día de la semana, Semana Santa[1]
PERNOCT_TOT	Tratamiento Aditivo por Shock	X-11	(2,0,0)(0,1,1)	Año bisiesto, Día de la semana, Semana Santa[1]

Nota: El método "Paramétrico Regular" emplea el motor paramétrico SEATS tras evaluar automáticamente la necesidad de transformaciones Box-Cox estabilizadoras. El método de "Intervención Exógena por Volatilidad" recurre al filtro X-11 para modelizar la variable *licit* con alta volatilidad, evitando así singularidades matriciales. Para el bloque de "Tratamiento Aditivo por Shock", se ha desactivado la transformación logarítmica predeterminada, imponiendo un esquema de descomposición aditiva pura sobre el motor X-11. Esta configuración permite el tratamiento matemático directo de las variables en nivelesa través de una trasnformación de desplazamiento ($y_t + 1$), absorbiendo los meses con actividad nula absoluta derivados del confinamiento de 2020 sin generar indeterminaciones.

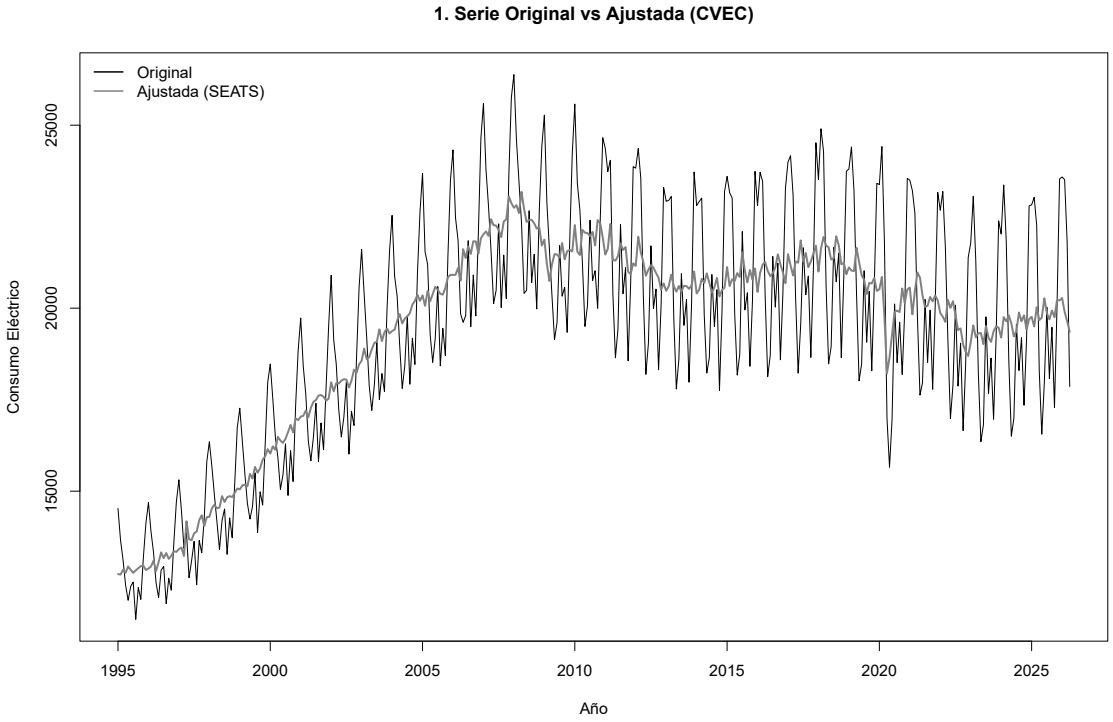


Figura 4.5: Serie original frente a serie ajustada (CVEC) para el Consumo Eléctrico (1995-2026).

.

la actividad agregada.

En segundo lugar, la adecuación del modelo generador se evalúa a través de la dinámica de sus residuos. La Figura 4.6 representa la Función de Autocorrelación (ACF) y la Función de Autocorrelación Parcial (PACF) de los residuos. A nivel visual, se constata que la práctica totalidad de los coeficientes carecen de estructura dinámica, situándose estrictamente dentro de las bandas de confianza asintóticas del 95 % (líneas discontinuas azules).

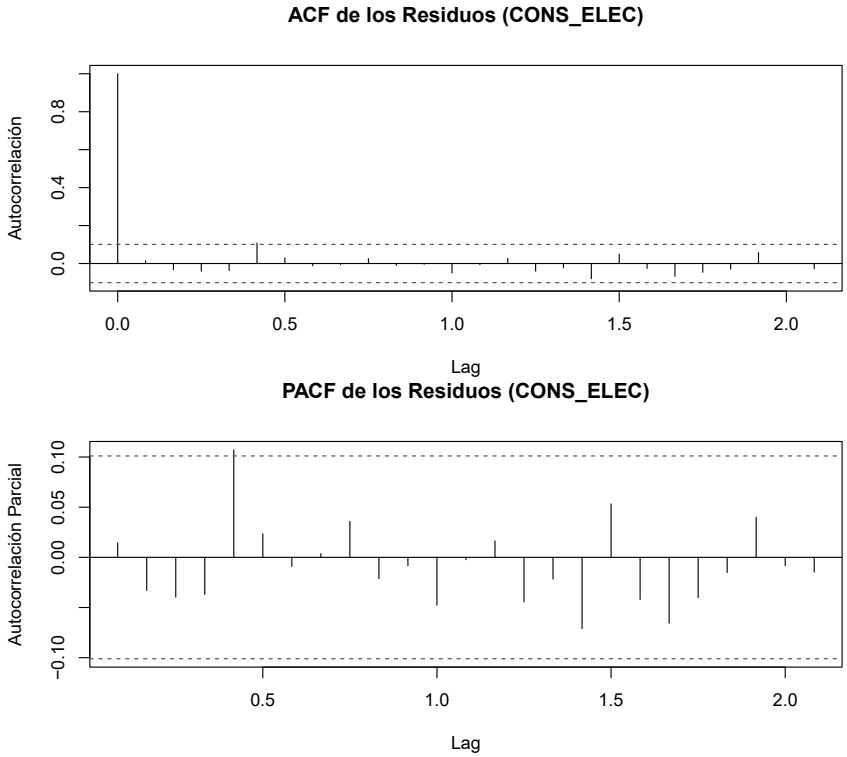


Figura 4.6: Función de Autocorrelación (ACF) y Función de Autocorrelación Parcial (PACF) de los residuos del modelo regARIMA.

.

La inspección se complementa y respalda mediante la validación a través de contrastes de hipótesis. Siguiendo el marco metodológico de Box-Jenkins para la diagnosis residual (véase la Sección A.5.3 del Apéndice A), se han sometido las innovaciones del modelo a una batería de contrastes de diagnóstico, cuyos resultados se sintetizan en la Tabla 4.4.

Como se infiere del cuadro resumen, los resultados validan las premisas fundamentales exigidas. Por un lado, la aplicación previa de transformaciones estabilizadoras en el algoritmo resolvió de forma endógena el requisito de *homocedasticidad*. Por otro lado, la estructura ARIMA $(1,1,1)(0,1,1)$ ha logrado filtrar la señal en su totalidad, devolviendo un error sin sesgo (media nula validada por el contraste t) y carente de inercia o memoria estocástica (incorrelación validada por el test de Ljung-Box). En conjunto, esto no arroja evidencia analítica en contra de que los residuos del modelo se comporten como ruido blanco. Finalmente, respecto a la premisa de *normalidad*, el test de Shapiro-Wilk detecta desviaciones empíricas respecto a la campana de Gauss. Este es un fenómeno asintótico habitual en paneles macroeconómicos de alta frecuencia temporal ($T = 376$), derivado de la presencia de *outliers* o colas pesadas.

Una vez se ha presentado un ejemplo representativo del flujo de trabajo que se ha seguido para cada una de la series del panel, a continuación, se vuelve a realizar una validación *ex-post* a través de

Tabla 4.4: Contrastes de diagnóstico paramétrico sobre los residuos del modelo ARIMA (1,1,1)(0,1,1) (CONS_ELEC)

.

Contraste	Estadístico	Conclusión
Test de Ljung-Box (Q)	17.10 ($p = 0{,}705$)	No se rechaza H_0. Ausencia de autocorrelación serial conjunta.
Media Nula (t-Student)	-0.585 ($p = 0{,}558$)	No se rechaza H_0. Ausencia de sesgo sistemático ($E[\varepsilon_t] = 0$).
Contraste QS	0.00 ($p = 1{,}000$)	No se rechaza H_0. Ausencia de estacionalidad residual en la serie (CVEC).
Test de Shapiro-Wilk (W)	0.959 ($p < 0{,}001$)	Se rechaza H_0. Ausencia de normalidad (desviación compensada vía QML).
Significatividad Paramétrica (z)	— ($p < 0{,}001$)	Se rechaza H_0. Totalidad de coeficientes estructurales altamente significativos.

Nota: H_0 denota la hipótesis nula de cada contraste y p el p.valor referente a cada estadístico. Nivel de significación de referencia: $\alpha = 0{,}05$.

los mismos contrastes individuales y la combinación ponderada con la función de R *isSeasonal*.

4.3.3. Estabilización de la varianza y transformación para la estacionariedad

Una vez garantizada la eliminación de los patrones estacionales y de calendario, las series temporales resultantes ($CVEC$) aún presentan propiedades que las inhabilitan para su introducción directa en el MFD. En concreto, la inmensa mayoría de las magnitudes macroeconómicas observadas exhiben varianzas no constantes a lo largo del tiempo (heterocedasticidad) y trayectorias tendenciales (no estacionariedad en media).

La exigencia de estacionariedad débil (Véase la Sección A.3 del Apéndice A para obtener una definición completa de este concepto) -integración de orden cero o $I(0)$- constituye un requisito matemático en el modelizado factorial.

Para solventar la heterocedasticidad, inherente al crecimiento exponencial a largo plazo de los volúmenes macroeconómicos, se ha aplicado una transformación logarítmica ($\log y_t$). Esta operación permite linealizar la trayectoria de crecimiento a largo plazo y estabilizar la varianza de las perturbaciones.

Una vez estabilizada la varianza, se ha procedido a la corrección de la no estacionariedad en media. En la literatura relacionada con el *Nowcasting*, el consenso metodológico establecido por Stock y Watson (2002a) dicta que el orden de diferenciación requerido depende intrínsecamente de la naturaleza económica del indicador. Bajo este paradigma, en el presente proyecto se ha adoptado una dicotomía para el tratamiento del panel:

- **Indicadores cualitativos (*Soft Data*):** Comprenden las encuestas de sentimiento y expectativas (como los índices PMI, el Indicador de Confianza del Consumidor o el Sentimiento Económico). Estas variables están acotadas por construcción en escalas cerradas (por ejemplo, oscilando entre 0 y 100) y, por definición teórica, carecen de tendencia a largo plazo, resultando imposible que alberguen una varianza infinita a largo plazo. En consecuencia, ingresan al modelo factorial estrictamente en niveles.

- **Indicadores cuantitativos (*Hard Data*):** Agrupan las variables de volumen, empleo e índices de actividad real. Estas series son procesos integrados de primer orden $I(1)$. Para inducir su estacionariedad, se ha aplicado una primera diferencia regular sobre la serie logarítmica ($\Delta \log y_t$), transformándolas matemáticamente en tasas de crecimiento mensual y trimestral, en el caso de la serie de Ocupados y el PIB.

Tabla 4.5: Diagnóstico de estacionalidad residual

.

		P-valores de los contrastes en la serie ajustada			
Variable	**¿Estacionalidad residual?**	**Kruskal-Wallis**	**Friedman**	**QS**	**Welch**
PARO	NO	0.999	0.700	0.000	0.940
IPI	NO	0.999	0.996	1.000	0.982
CEM	NO	0.941	0.877	1.000	0.994
CONS_ELEC	NO	1.000	1.000	1.000	1.000
ICM	NO	0.984	0.799	1.000	0.791
AFIL_SS	NO	0.000	0.000	1.000	1.000
MAT_CARGA	NO	0.855	0.737	1.000	1.000
VIS	NO	0.867	0.923	1.000	1.000
LICIT	NO	0.923	0.987	1.000	0.741
ICN_SERV	NO	0.999	0.976	1.000	0.912
MAT_TUR	NO	0.969	0.772	1.000	1.000
EXP_BIENES	NO	0.993	0.997	1.000	1.000
IMP_BIENES	NO	0.980	0.951	1.000	1.000
CP_VIV	NO	0.794	0.903	1.000	1.000
PASAJ_AER	NO	0.453	0.501	1.000	1.000
GAS_TUR	NO	0.574	0.462	1.000	0.789
TURIST_INT	NO	0.274	0.244	1.000	0.272
PERNOCT_TOT	NO	0.254	0.278	1.000	0.697
TRAF_PORT	NO	1.000	1.000	1.000	1.000
ICC_CONS	NO	0.147	0.145	1.000	0.104
ICC	NO	0.901	0.894	1.000	0.960
ISE	NO	0.955	0.940	1.000	0.890
IPI_CONS	NO	0.969	0.977	0.887	0.728
GASOLINA+DIESEL	NO	0.349	0.210	1.000	1.000
GAS	NO	1.000	0.998	1.000	1.000
FIN_FAM_NUEVAS	NO	0.143	0.134	1.000	0.990
VGE_TOT	NO	0.993	0.999	1.000	0.933
VGE_SAL	NO	0.962	0.893	0.400	0.932
PMI_SERV	NO	0.717	0.926	1.000	0.401
PMI_MANU	NO	0.373	0.329	1.000	0.328
OCU	NO	0.959	0.946	0.147	0.836
PIB	NO	0.945	0.632	1.000	0.430

Para validar este marco de transformaciones, se ha sometido el panel completo al contraste de raíz unitaria de Dickey-Fuller Aumentado (ADF) (véase la Sección A.5.1 del Apéndice A) a través de la función *adf.test* perteneciente a la librería de R *tseries*. Este test evalúa como hipótesis nula (H_0) que la serie generadora es un proceso no estacionario. La Tabla 4.6 expone los p-valores resultantes tanto en los niveles de origen como tras el preprocesado propuesto.

Los resultados ilustran las limitaciones prácticas de los contrastes de raíz unitaria en muestra finita frente a variables de elevada persistencia. Por un lado, variables basadas en encuestas cualitativas y de expectativas (`ICC`, `ICC_CONS` e `ISE`) no logran rechazar la hipótesis nula de raíz unitaria en niveles al 5 %. Este comportamiento anómalo se debe a que la dinámica a corto plazo de estos indicadores responde a un proceso con una raíz cercana a la unidad. Desde una perspectiva teórica, estas variables cualitativas son intrínsecamente estacionarias $I(0)$ debido a que se encuentran acotadas por construcción dentro de un rango finito (por ejemplo, los índices PMI están restringidos al intervalo $[0, 100]$ y los saldos de opinión a $[-100, 100]$). Dado que un proceso integrado $I(1)$ requiere que su varianza temporal tienda a infinito, la existencia de estos límites físicos imposibilita matemáticamente la presencia de una verdadera raíz unitaria. Sin embargo, la alta persistencia temporal de las expectativas macroeconómicas merma la potencia estadística del test ADF en muestras finitas, provocando que el contraste confunda la dinámica de corto plazo con la de un proceso integrado (Error Tipo II)[4]. Este fenómeno justifica formalmente la decisión metodológica de incorporar los indicadores de *soft data* directamente en niveles, preservando intacta tanto su interpretación económica original como su capacidad para capturar los puntos de giro del ciclo.

Continuando con el caso de ejemplo del Consumo de Energía Eléctrica (`CONS_ELEC`), la serie corregida de estacionalidad y calendario (CVEC) obtenida en la fase anterior se clasifica estrictamente dentro del bloque de indicadores cuantitativos de actividad real (*Hard Data*). Al tratarse de una magnitud de volumen expuesta al crecimiento de largo plazo, la serie presenta una marcada heterocedasticidad dependiente del nivel, como se observó en la Figura 4.5. Para mitigar esta inestabilidad en la varianza se le aplica, en primer lugar, la transformación logarítmica de la familia de Box-Cox ($\lambda = 0$).

En segundo lugar, para corregir la trayectoria tendencial no estacionaria en media, se ejecuta una primera diferencia regular sobre el componente linealizado. De este modo, la transformación final de la serie se expresa formalmente como la tasa de crecimiento intermensual de la demanda eléctrica:

$$\Delta \ln(\text{CONS\_ELEC}_t) = \ln(\text{CONS\_ELEC}_t) - \ln(\text{CONS\_ELEC}_{t-1})$$

La validez estadística de este procedimiento se constata al examinar los resultados del contraste de raíz unitaria recogidos en la Tabla 4.6. En su estado original en niveles (CVEC), el test de Dickey-Fuller Aumentado (ADF) arroja un p-valor de 0,833, lo que impide rechazar la hipótesis nula de no estacionariedad debido a su elevada persistencia. Por el contrario, tras aplicar la primera diferencia logarítmica regular ($\Delta \log$), el contraste ADF devuelve un p-valor inferior a 0,010. Este resultado permite rechazar la H0 sobre presencia de una raíz unitaria al nivel de significación del 5 %, garantizando que la serie transformada es un proceso estacionario integrado de orden cero, $I(0)$, apto para la modelización factorial. A través de la evidencia gráfica de la Figura 4.7 se observa el éxito de esta transformación sobre el Consumo de Energía Eléctrica (`CONS_ELEC`). Tras aplicar el filtro de diferencias, la serie pierde su inercia tendencial y pasa a moverse en torno al cero. Asimismo, la transformación logarítmica previa estabiliza la banda de dispersión, corrigiendo el patrón de heterocedasticidad creciente que caracterizaba a los niveles de origen.

[4]En teoría de contraste de hipótesis, el **error de tipo II** (o falso negativo) se define como la aceptación de una hipótesis nula (H_0) que en realidad es falsa. En el contexto específico del test de Dickey-Fuller Aumentado (ADF), donde la H_0 postula la existencia de una raíz unitaria (no estacionariedad), este error se materializa cuando el contraste concluye que la serie es integrada de orden uno, $I(1)$, a pesar de ser teóricamente estacionaria, $I(0)$. Este fallo es el resultado directo de la baja potencia del test en muestras finitas para discriminar raíces autorregresivas muy cercanas a la unidad ($\rho \approx 0{,}98$).

Tabla 4.6: Contraste de Raíz Unitaria (Dickey-Fuller Aumentado) previo y posterior a la transformación .

	P-valor (Test ADF)		
Variable	**En niveles (CVEC)**	**Tras transformación**	**Conclusión Final**
PARO	0.886	0.056	$I(0)$ en $\Delta \log$ (al 10 %)
IPI	0.486	$< 0{,}010$	$I(0)$ en $\Delta \log$
CEM	0.706	$< 0{,}010$	$I(0)$ en $\Delta \log$
CONS_ELEC	0.833	$< 0{,}010$	$I(0)$ en $\Delta \log$
ICM	0.942	$< 0{,}010$	$I(0)$ en $\Delta \log$
AFIL_SS	0.930	$< 0{,}010$	$I(0)$ en $\Delta \log$
MAT_CARGA	0.664	$< 0{,}010$	$I(0)$ en $\Delta \log$
VIS	0.736	$< 0{,}010$	$I(0)$ en $\Delta \log$
LICIT	0.528	$< 0{,}010$	$I(0)$ en $\Delta \log$
ICN_SERV	0.943	$< 0{,}010$	$I(0)$ en $\Delta \log$
MAT_TUR	0.519	$< 0{,}010$	$I(0)$ en $\Delta \log$
EXP_BIENES	0.394	$< 0{,}010$	$I(0)$ en $\Delta \log$
IMP_BIENES	0.310	$< 0{,}010$	$I(0)$ en $\Delta \log$
CP_VIV	0.038	$< 0{,}010$	$I(0)$ en $\Delta \log$
PASAJ_AER	0.182	$< 0{,}010$	$I(0)$ en $\Delta \log$
GAS_TUR	0.664	0.078	$I(0)$ en $\Delta \log$ (al 10 %)
TURIST_INT	0.600	$< 0{,}010$	$I(0)$ en $\Delta \log$
PERNOCT_TOT	0.125	$< 0{,}010$	$I(0)$ en $\Delta \log$
TRAF_PORT	0.402	$< 0{,}010$	$I(0)$ en $\Delta \log$
ICC_CONS	0.782	*No aplica*	$I(0)$ teórica (Niveles)
ICC	0.099	*No aplica*	$I(0)$ teórica (Niveles)
ISE	0.102	*No aplica*	$I(0)$ teórica (Niveles)
IPI_CONS	0.865	$< 0{,}010$	$I(0)$ en $\Delta \log$
GASOLINA+DIESEL	0.418	$< 0{,}010$	$I(0)$ en $\Delta \log$
GAS	0.130	$< 0{,}010$	$I(0)$ en $\Delta \log$
FIN_FAM_NUEVAS	0.649	$< 0{,}010$	$I(0)$ en $\Delta \log$
VGE_TOT	0.702	$< 0{,}010$	$I(0)$ en $\Delta \log$
VGE_SAL	0.972	$< 0{,}010$	$I(0)$ en $\Delta \log$
PMI_SERV	0.046	*No aplica*	$I(0)$ en Niveles
PMI_MANU	0.045	*No aplica*	$I(0)$ en Niveles
OCU	0.835	$< 0{,}010$	$I(0)$ en $\Delta \log$
PIB	0.513	$< 0{,}010$	$I(0)$ en $\Delta \log$

Nota: H_0: La serie es no estacionaria (raíz unitaria).

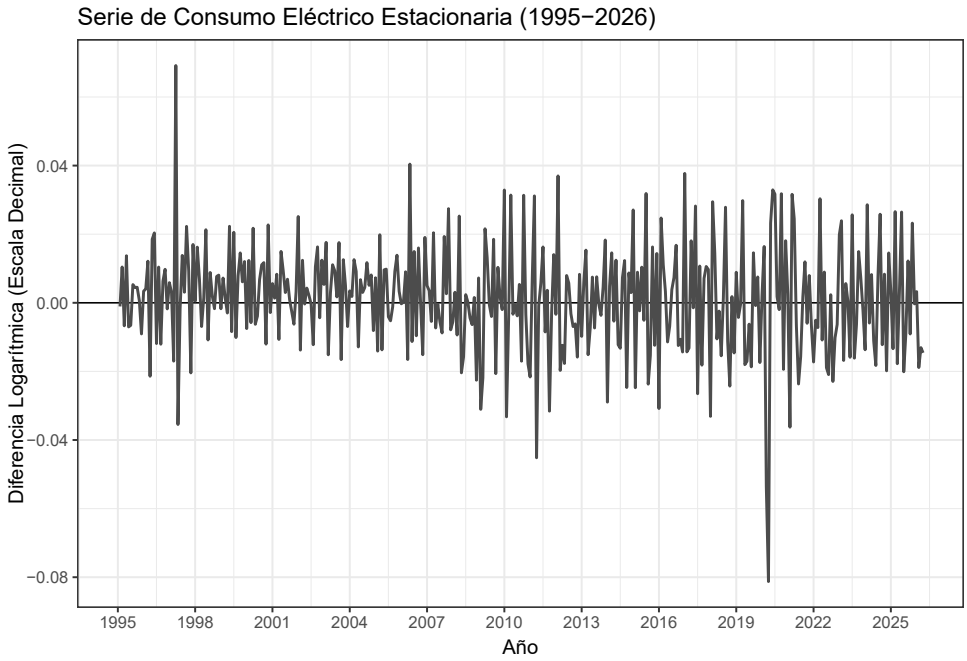


Figura 4.7: Tasa de variación mensual estacionaria ($\Delta \log$) del Consumo de Energía Eléctrica (1995-2026).

.

4.3.4. Estandarización de las variables

El último paso del pre-procesado antes de meter el panel de datos en el MFD es la homogeneización de sus escalas. Tras el ajuste estacional y las transformaciones orientadas a garantizar la estacionariedad, la matriz de entrada sigue estando compuesta por magnitudes estadísticamente heterogéneas. Mientras que las variables transformadas mediante diferencias logarítmicas se expresan en tanto por uno con órdenes de magnitud decimales (y varianzas extremadamente reducidas), los índices de difusión como el PMI operan en niveles acotados en torno a los 50 puntos, exhibiendo una dispersión numéricamente muy superior.

Desde un punto de vista puramente matemático, la estimación de los factores dinámicos es un proceso extremadamente sensible a la varianza original de los datos. Si las variables se introdujeran en su métrica natural, el algoritmo de extracción factorial otorgaría un peso artificialmente desproporcionado a aquellas series con mayor amplitud numérica, silenciando por completo la señal macroeconómica de los indicadores con varianzas más pequeñas, independientemente de su verdadera relevancia cíclica.

Se aplica una tipificación estadística estándar (o *Z-score*) a cada serie individual, garantizando que todas compitan en igualdad de condiciones en la extracción de la varianza común. La transformación matemática se define como:

$$z_{i,t} = \frac{x_{i,t} - \hat{\mu}_i}{\hat{\sigma}_i},$$

donde para cada variable i en el instante t, la observación de la serie estacionaria $x_{i,t}$ se centra restándole su media muestral ($\hat{\mu}_i$) y se escala dividiéndola por su desviación típica muestral ($\hat{\sigma}_i$). Como resultado, se obtiene una nueva matriz de variables estandarizadas ($\mathbf{Z}$) donde todas las series presentan por construcción una media nula ($\hat{\mu} = 0$) y una varianza unitaria ($\hat{\sigma}^2 = 1$).

Para cerrar este capítulo se presenta el caso de ejemplo de la demanda eléctrica, la variable

$\Delta \log(\text{CONS\_ELEC}_t)$ se somete a este centrado y escalado. El resultado de esta última transformación matemática se proyecta visualmente en la Figura 4.8:

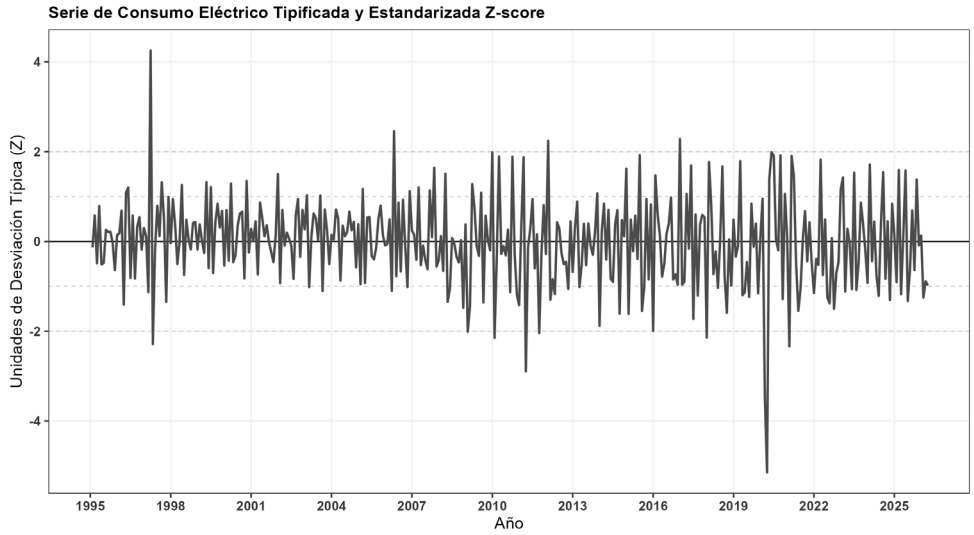


Figura 4.8: Serie tipificada y estandarizada Z-score del Consumo de Energía Eléctrica (1995-2026).

.

Como se evidencia en la Figura 4.8, la trayectoria temporal de la serie y la eliminación del patrón estacional son formalmente idénticas a las obtenidas en la consecución de la estacionariedad. No obstante, el impacto de la estandarización reside en su escala métrica: el eje de ordenadas ya no refleja la magnitud decimal de una tasa de crecimiento mensual, sino que se expresa estrictamente en unidades de desviación típica.

De esta manera, la variable presenta ahora una media aritmética de cero y una varianza unitaria por construcción ($z_{it} \sim (0, 1)$). Este paso garantiza que el Consumo de Energía Eléctrica compita en igualdad de condiciones en la sección transversal frente a variables cualitativas como los índices PMI (cuya dispersión original en niveles es muy superior), previniendo sesgos en la ponderación de pesos y asegurando que el MFD extraiga de forma equilibrada la señal común del ciclo.

La exhaustiva batería de transformaciones descrita a lo largo de este capítulo -desde la depuración de la estacionalidad y los efectos de calendario (con especial atención a los ceros estructurales provocados por la crisis de la COVID-19), hasta la transformaciones orientadas a garantizar la estacionariedad y la estandarización final se lleva cabo principalmente porque los Modelos Factoriales Dinámicos operan bajo el paradigma del aprendizaje no supervisado; el algoritmo extrae a ciegas la máxima varianza compartida en el panel de datos. Si se introdujeran series brutas, el factor latente estimado capturaría simplemente la estacionalidad común, la inflación nominal o las disrupciones de calendario, arrojando un resultado espurio. Con el procesado de serie realizado en este capítulo, se consigue que el factor dinámico que se extraerá en las etapas posteriores representará, de forma pura y aislada, el ciclo macroeconómico.

Capítulo 5

Selección y Especificación del Modelo Factorial Dinámico

Durante el presente capítulo se aborda la fase de construcción, criba y especificación del modelo final. De acuerdo con el marco metodológico de las aproximaciones factoriales en alta dimensión, la transición desde un panel extenso de indicadores sectoriales hacia un espacio latente de dimensión reducida requiere justificar secuencialmente tanto las decisiones de software como el tratamiento de la volatilidad histórica, la arquitectura de validación fuera de la muestra y la determinación de los parámetros estructurales del sistema.

5.1. Implementación del modelo: El paquete `dfms` en R

La estimación del Modelo Factorial Dinámico (MFD) formalizado en este proyecto se ha llevado a cabo con el software estadístico R a través de la librería especializada `dfms` creada por Sebastian Krantz y Rytis Bagdziunas (Krantz y Bagdziunas, 2023). Este paquete implementa algoritmos optimizados para la extracción y estimación de vectores de estado latentes mediante la recursión del Filtro de Kalman y el algoritmo de Maximización-Expectación (EM). La arquitectura del software toma como pilar y especificación de referencia el algoritmo de Bańbura y Modugno (2014), el método seleccionado para implantarlo en el proyecto de los presentados en el Capítulo 3. Cuando se usa la función principal del paquete, la función *DFM()*, el algoritmo ejecuta una estandarización automática sobre todas las series (media cero y varianza unitaria). Por lo tanto, a la función le llegan los datos CVEC e I(0) ya que los estandariza internamente. Una de las mayores ventajas del paquete `dfms` es que prescinde de la necesidad de aplicar técnicas exógenas de interpolación lineal manual sobre los baches informativos. La heterogeneidad de los calendarios de publicación institucional introducen de forma recurrente valores perdidos de carácter temporal, los cuales son asimilados directamente por el Espacio de Estados de forma endógena.

Ante la presencia de un valor clasificado como ausente (`NA`) en el instante t para una variable determinada, el algoritmo altera dinámicamente la ecuación de medida anulando los componentes correlativos de la Ganancia de Kalman ($K_t = \mathbf{0}$). Mediante este mecanismo, el filtro suspende la fase de actualización observacional para esa serie y proyecta su trayectoria basándose exclusivamente en la dinámica autorregresiva de la ecuación de transición. En escenarios de ausencia prolongada de información —característicos de las variables de baja frecuencia intermitente—, este procedimiento opera como una interpolación que converge asintóticamente hacia la media incondicional del proceso, blindando la matriz de covarianzas frente a distorsiones artificiales y asegurando la estabilidad paramétrica global.

5.1.1. Configuración de la función `DFM`

La estimación del modelo se ejecuta a través de la función principal `DFM()`, se requiere de la selección de un conjunto de hiperparámetros estructurales:

- `r`: Define de forma explícita el número de factores latentes (r) que el algoritmo debe extraer del panel de datos.
- `p`: Establece el orden del proceso autorregresivo (p) que rige la dinámica temporal de la ecuación de transición del estado común.
- `idio.ar1`: Parámetro lógico (`TRUE/FALSE`) que introduce una estructura autorregresiva de primer orden (AR(1)) en las perturbaciones específicas de cada serie, permitiendo transicionar desde la rigidez del modelo exacto hacia la flexibilidad de la representación de Vector de Estado Ampliado.
- `em.method`: Comando de selección que permite conmutar el algoritmo de optimización numérica, pudiendo optar por la metodología de Máxima Verosimilitud Cuasi-Exacta de Bańbura y Modugno (2014) (`"BM"`) para realizar actualizaciones paramétricas fila por fila bajo la presencia de datos faltantes.
- `quarterly.vars`: Vector de indexación que identifica a los indicadores del panel que poseen una frecuencia de observación trimestral, activando de forma automática las restricciones de agregación linealizada dentro de la matriz de medida.
- `max.missing`: Cota de tolerancia que restringe la fracción máxima admisible de valores ausentes por fila, garantizando la estabilidad y consistencia de los Componentes Principales utilizados en la fase de inicialización paramétrica del algoritmo.

Una vez finalizadas las iteraciones del algoritmo EM y alcanzado el criterio de convergencia para el vector de parámetros estructurales, el paquete ofrece funciones para la inferencia de las variables inobservables. Mediante la aplicación de la función general de ajuste bajo el argumento de control `use.full.state = TRUE`, el entorno de ejecución despliega el algoritmo del Suavizador de Kalman.

A diferencia del filtrado tradicional en tiempo real, el suavizador realiza una doble recursión secuencial (un recorrido cronológico hacia adelante seguido de un retroceso dinámico hacia atrás) sobre la totalidad de la muestra histórica disponible (T). Este procedimiento refina las estimaciones del estado latente al condicionarlas a la información completa del panel, proporcionando trayectorias suavizadas tanto para el factor común como para los ruidos aleatorios autorregresivos, minimizando el error cuadrático medio intramuestral y permitiendo la posterior Descomposición de Noticias (*News*).

5.2. Tratamiento del período COVID-19

La irrupción de la crisis sanitaria de la COVID-19 a principios del año 2020 no constituyó una recesión ordinaria originada por desequilibrios financieros, como la crisis de 2007-08 o la de inicios del siglo XXI de las *.com*, sino un shock exógeno de gran magnitud (confinamientos estrictos y suspensión forzosa de actividades no esenciales). Desde una perspectiva macroeconómica, esta paralización de los flujos de producción, consumo y empleo generando outliers de gran magnitud que alteran las series de tiempo macroeconómicas. El impacto de la pandemia introdujo observaciones atípicas de de gran magnitud que alteran las relaciones históricas entre las variables económicas. A nivel matemático, esta distorsión afecta directamente a la estimación lineal en el espacio de estados. El Filtro de Kalman aplicado a un MFD asume que las perturbaciones ε_t y η_t siguen una distribución normal y tienen varianza constante (homocedasticidad). Sin embargo, la volatilidad extrema durante el confinamiento puede provocar romper el supuesto de varianza constante e incrementa de forma desproporcionada las matrices de covarianza de los errores. Al alterarse la relación señal-ruido, la ganancia de Kalman (K_t) se desajusta. Como consecuencia, el algoritmo puede confundir un *shock* externo y transitorio con un

cambio permanente en la tendencia del factor común y empeorar la capacidad predictiva del modelo. Por tanto, la crisis de 2020 no se comporta como una recesión ordinaria, sino como una perturbación externa que introduce una volatilidad extrema y altera las relaciones estadísticas tradicionales entre las variables del panel. Esta sección justifica, en primer lugar, la necesidad o no de una intervención sobre dicho periodo; en segundo lugar, delimita una ventana temporal de tratamiento (5.2.1); y, finalmente, se describen las tres estrategias evaluadas sobre el panel de datos 5.2.2. La literatura reciente sobre el MFD documenta que el comovimiento de los indicadores macroeconómicos durante la crisis sanitaria difieren estructuralmente de los observados en las recesiones precedentes. Maroz, Stock, y Watson (2021) demuestran que la crisis de 2020 pudo afectar fuertemente a los supuestos de normalidad y homocedasticidad sobre los que se apoyan los algoritmos de filtrado lineal como el Filtro de Kalman que se presentan en la Sección 3.2.1. Para ilustrar esta anomalía, los autores comparan la distribución de los errores de predicción fuera de la muestra en unidades de desviación típica ($\hat{\sigma}$): mientras que en la Gran Recesión (2007-2009) la mediana de la distribución de los indicadores se situó en $-1{,}7\hat{\sigma}$ (y el percentil 25 en $-2{,}8\hat{\sigma}$), durante la caída sufrida por la COVID-19 la mediana se hundió hasta un $-14{,}8\hat{\sigma}$. Algunos valores detectados en las colas de la distribución alcanzaron cotas sin precedentes históricos, registrándose desviaciones extremas de hasta $-275\hat{\sigma}$ en indicadores específicos como el empleo del sector de la hostelería. Finalmente, a través de la estimación de un factor latente de error estandarizado se pudo identificar el periodo de mayor desviación desde marzo hasta julio de 2020. En cuanto al caso del modelo creado y actualizado por el banco de España Arencibia Pareja et al. (2024), que ya se citó en el capítulo anterior, documentan un diagnóstico análogo para su MFD: la inclusión sin corrección del periodo de la pandemia en la muestra de estimación incrementa los errores absolutos de previsión, triplica la volatilidad del factor latente común e invierte el signo de las cargas factoriales (λ_i) de algunos indicadores clave.Con base en esta evidencia, en el presente trabajo se adopta esta metodología propuesta por Maroz et al. (2021) para la identificación de observaciones a intervenir a través del factor latente de error estandarizado.

5.2.1. Identificación de la ventana temporal del *shock*

Para fundamentar empíricamente la duración de esta ventana, se replica sobre un subconjunto reducido de indicadores de *hard data* con gran representatividad de la economía española [1] el siguiente procedimiento:

1. Las series se estandarizan utilizando exclusivamente la media y la desviación típica calculadas sobre el periodo de estabilidad pre-pandemia (hasta febrero de 2020).

2. Sobre el subpanel estandarizado pre-pandemia se aplica un PCA y se extrae la primera componente principal, cuyo vector de cargas w_1 resume la estructura histórica de comovimiento,el análogo a las cargas factoriales Λ del MFD.

3. Estas cargas históricas se recalculan sobre la muestra completa (pre-pandemia y pandemia) para obtener un factor extrapolado $\hat{f}_t = X_t w_1$ y su correspondiente reconstrucción ajustada $\hat{X}_t = \hat{f}_t w_1'$, en un ejercicio equivalente a la extrapolación OLS del MFD sobre el periodo COVID descrita por Maroz et al. (2021)[2].

4. El residuo de esta proyección, $u_t = X_t - \hat{X}_t$, recoge la parte de la varianza del panel que la estructura de comovimiento pre-pandemia es incapaz de explicar en cada instante t.

[1] Índice de Producción Industrial, Indicador de Cifra de Negocios de Servicios, Ventas de Grandes Empresas Totales , Exportaciones e Importaciones de bienes y Tráfico de pasajeros en Aeropuertos.

[2] A diferencia de estos autores, que contrastan un estimador OLS frente a un estimador WLS (Weigthted Least Squares) robusto a valores atípicos mediante una mixtura de normales, la extrapolación aquí empleada es unicamente OLS (vía componentes principales); resulta suficiente para el objetivo de esta subsección, que es exclusivamente delimitar la ventana temporal de tratamiento, y no la estimación final de las cargas factoriales del modelo.

5. Finalmente, se extrae la primera componente principal de la matriz de residuos u_t -un *factor de error estandarizado*- cuya trayectoria durante el periodo 2019-2024 se puede observar en la Figura 5.1.

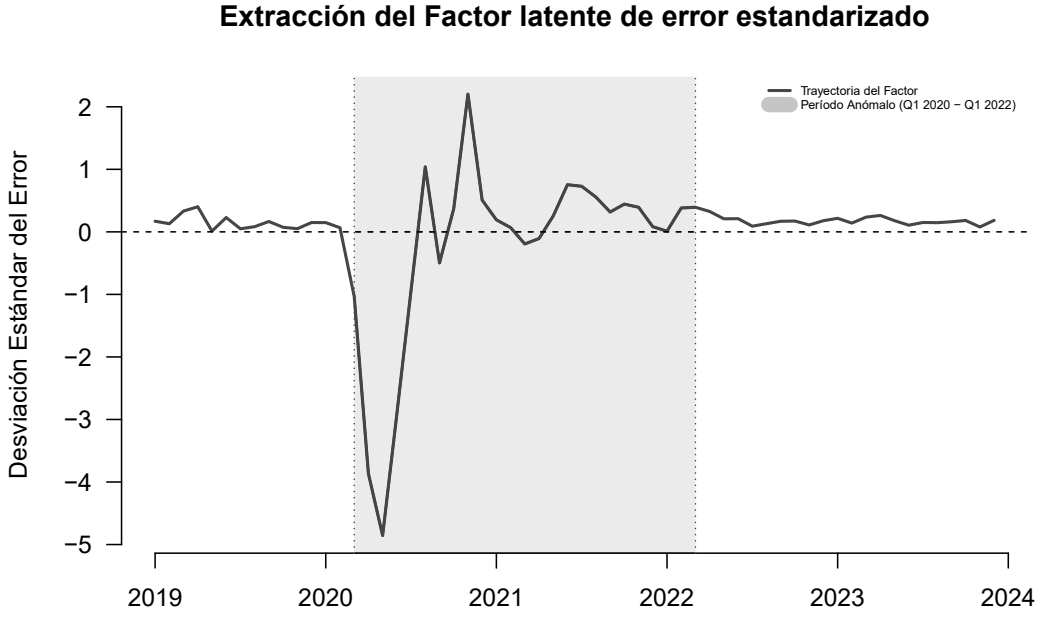


Figura 5.1: Trayectoria del factor latente de error estandarizado, 2019–2023

.

Tabla 5.1: Valores numéricos del factor latente de error estandarizado ($\hat{\sigma}$).

.

Año	**Ene**	**Feb**	**Mar**	**Abr**	**May**	**Jun**
2019	0.26	−0,51	3.44	4.79	−2,82	1.42
2020	−0,17	−1,76	−23,19	−78,53	**−97,86**	−59,93
2021	0.71	−1,72	−6,82	−5,14	1.92	11.67
2022	−2,84	4.40	4.63	3.32	1.02	1.05
2023	1.17	−0,31	1.54	2.07	0.44	−0,97
Año	**Jul**	**Ago**	**Sep**	**Oct**	**Nov**	**Dic**
2019	−2,10	−1,43	0.18	−1,63	−2,07	−0,14
2020	−21,28	17.24	−12,74	4.15	**39,94**	6.86
2021	11.18	7.80	3.15	5.61	4.63	−1,45
2022	−1,29	−0,50	0.26	0.32	−0,90	0.44
2023	−0,12	−0,19	0.10	0.51	−1,53	0.53

Nota: Valores expresados en unidades de desviación típica ($\hat{\sigma}$), reescaladas sobre la ventana 2019-2023 para su representación gráfica. Las líneas verticales discontinuas en la Figura 5.1 delimitan la ventana de tratamiento máxima identificada (T1:2020–T1:2022).

La Figura 5.1 permite distinguir tres fases diferenciadas. En primer lugar, la línea de base de 2019 se mantiene, en términos generales, sobre la media teórica de 0, si bien se observan valores puntuales

de hasta $4{,}79\hat{\sigma}$ (abril de 2019); esta amplitud, mayor de lo se espera bajo un periodo de estabilidad, es atribuible al tamaño de la ventana muestral disponible para la estimación de σ_i^{pre}. Esta pérdida de historia es debido a la necesidad de tener un panel de variables balanceado, es decir, sin datos faltantes, para poder estimar el PCA; siendo una limitación que se ha encontrado producto de la asincronía y heterogeneidad de las series seleccionadas.[3]

En segundo lugar, a partir de marzo de 2020 el factor de error experimenta una ruptura, alcanzando su valor mínimo en mayo de 2020 ($-97{,}86\hat{\sigma}$), tras una fuerte caída ya desde marzo ($-23{,}19\hat{\sigma}$) y abril ($-78{,}53\hat{\sigma}$). La recuperación posterior no es monótona: se observa un repunte en agosto de 2020 ($+17{,}24\hat{\sigma}$), seguido de una nueva caída en septiembre ($-12{,}74\hat{\sigma}$) y, posteriormente, un segundo pico en noviembre de 2020 ($+39{,}94\hat{\sigma}$), que coincide temporalmente con la declaración del segundo estado de alarma en España (25 de octubre de 2020). Durante 2021 el factor se mantiene lejos de la base en 0, con valores que se mueven en torno a $+11\hat{\sigma}$ entre junio y julio.

Finalmente, a partir de 2022 el factor de error se reubica en un rango más moderado, con algunas oscilaciones de hasta $\pm 4{,}6\hat{\sigma}$ (febrero-marzo de 2022) que se reducen progresivamente a lo largo del año, aproximándose de forma asintótica a la amplitud observada en la línea de base de 2019.

Los resultados obtenidos proporcionan una justificación para la delimitación de la ventana de tratamiento, y resulta consistente con la estrategia adoptada por organismos de información macroeconómica, como el Banco de España y la AIReF, anteriormente citados sus trabajos donde lo realizaban sobre sus MFD para el *nowcasting*. Se concluye que el efecto sobre la estructura de covarianzas se diluye de manera sustancial a partir del primer trimestre de 2022, aunque no de forma instantánea, dado que el factor conserva cierta amplitud residual (en torno a $0{,}3$–$0{,}4\hat{\sigma}$) durante buena parte de dicho ejercicio. En conclusión, el algoritmo de tratamiento de valores atípicos descrito en la Sección 5.2.2 permite identificar 2 periodos delimitados para el tratamiento de la COVID-19:

1. Desde marzo hasta julio de 2020 , en adelante periodo-1-covid, por ser el periodo con el mayor pico desviaciones típicas del error del periodo de análisis.

2. Desde marzo de 2020 hasta marzo de 2022, en adelante periodo-2-covid, por ser el periodo desde el inicio de la caída del factor del error hasta su convergencia final en marzo de 2022 a valores cercanos a 0 de forma continuada.

5.2.2. Estrategias de tratamiento evaluadas

Una vez identificadas las ventanas temporales con las que se pretende intervenir la perturbación creada por la COVID-19, a continuación, se presentan las distintas pruebas o estrategias que se han seguido:

1. **Estimación sobre el panel bruto (línea base sin tratamiento).** El modelo se estima incluyendo la totalidad de las observaciones de 2020 sin ninguna corrección, y actúa como grupo de control para dimensionar el impacto de la perturbación sobre las cargas factoriales y la capacidad predictiva fuera de la muestra.

2. **Imputación mediante el Filtro de Kalman (valores ausentes)** Esta estrategia consiste en codificar como valores ausentes (*NA*) la totalidad de las observaciones correspondientes a la ventana anómala delimitada en la Sección 5.2.1. El fundamento de este procedimiento se apoya en el estimador de máxima verosimilitud de Bańbura y Modugno (2014) bajo el algoritmo EM, adaptando el sistema en el Espacio de Estados definido en el Capítulo 3. Para preservar la coherencia con la formalización matemática del filtro general, adoptamos la equivalencia donde el vector observable $\mathbf{y}_t$ corresponde al panel $\mathbf{X}_t$, el vector de estados $\mathbf{a}_t$ representa a los factores latentes $\mathbf{F}_t$, la matriz de medida $\mathbf{Z}_t$ equivale a las cargas factoriales $\mathbf{\Lambda}$, y la covarianza del error $\mathbf{H}_t$ se denota como $\mathbf{R}_t$:

[3] Con una muestra pre-COVID pequeña, la desviación típica estimada está sujeta a un posible error de muestreo, lo que puede inflar los cocientes de tipificación.

$$\mathbf{X}_t = \mathbf{\Lambda}\mathbf{F}_t + \mathbf{e}_t, \quad \mathbf{e}_t \sim N(\mathbf{0}, \mathbf{R}_t)$$

.

$$\mathbf{F}_t = \mathbf{\Phi}\mathbf{F}_{t-1} + \mathbf{G}\boldsymbol{\eta}_t, \quad \boldsymbol{\eta}_t \sim N(\mathbf{0}, \mathbf{Q}_t).$$

Ante la presencia de vectores desbalanceados en el instante t, el algoritmo implementa la matriz de selección determinista $\mathbf{W}_t$ introducida en el Corolario 3.1. Mediante esta transformación, el sistema se reduce dinámicamente a $\mathbf{X}_t^* = \mathbf{W}_t\mathbf{X}_t$, $\mathbf{\Lambda}_t^* = \mathbf{W}_t\mathbf{\Lambda}$ y $\mathbf{R}_t^* = \mathbf{W}_t\mathbf{R}_t\mathbf{W}_t'$, parametrizando la Ganancia de Kalman como:

$$\mathbf{K}_t^* = \mathbf{P}_{t|t-1}\mathbf{\Lambda}_t^{*\prime}\left(\mathbf{\Lambda}_t^*\mathbf{P}_{t|t-1}\mathbf{\Lambda}_t^{*\prime} + \mathbf{R}_t^*\right)^{-1}.$$

De este modo, la actualización del estado se computa sobre el subespacio disponible: $\mathbf{F}_{t|t} = \mathbf{F}_{t|t-1} + \mathbf{K}_t^*\left(\mathbf{X}_t^* - \mathbf{\Lambda}_t^*\mathbf{F}_{t|t-1}\right)$. En el caso límite objeto de este tratamiento, donde la totalidad del vector transversal se codifica como ausente en el mes t, la matriz de selección se anula ($\mathbf{W}_t = \mathbf{0}$), lo que suspende de forma automática el proceso de actualización al cancelarse la Ganancia de Kalman ($\mathbf{K}_t^* = \mathbf{0}$).

Bajo este escenario, el algoritmo interrumpe la fase de corrección y se reduce estrictamente a la fase de predicción ($\mathbf{F}_{t|t} = \mathbf{F}_{t|t-1}$). Al estar el modelo calibrado en log-diferencias estacionarias, la estabilidad estricta de la matriz de transición $\mathbf{\Phi}$ induce un decaimiento asintótico de los factores hacia su media inobservable condicional (cero). El suavizador de Kalman opera así una interpolación basada exclusivamente en la inercia autorregresiva estimada en el período histórico pre-pandemia, evitando de forma efectiva que la varianza extrema de 2020 contamine la reestimación de los parámetros estructurales en la etapa M del algoritmo EM.

3. **Imputación univariante (proyecciones ARIMA).** Como alternativa exógena al procedimiento anterior, los registros de la ventana de tratamiento delimitada en la Sección 5.2.1 se sustituyen, con carácter previo a la estimación del MFD, por las proyecciones puntuales derivadas de modelos autorregresivos integrados de medias móviles (ARIMA). Estos modelos se calibran serie a serie utilizando exclusivamente el histórico de información disponible hasta febrero de 2020.

   Desde un punto de vista analítico, esta estrategia proporciona al algoritmo multivariante una predicción condicional univariante de referencia para el período intervenido. Al suministrar un vector de datos completo, la matriz de selección definida en la ecuación de medida se preserva constante como una matriz identidad ($\mathbf{W}_t = \mathbf{I}$) a lo largo de toda la muestra. Esto evita la alteración dinámica de la dimensionalidad del sistema y la consecuente suspensión de la etapa de actualización del filtro. La sustitución del confinamiento por la predicción condicional univariante mitiga la inflación de la matriz de covarianzas muestral ($\mathbf{R}_t$), garantizando que el algoritmo EM opere sobre un panel balanceado donde la Ganancia de Kalman ($\mathbf{K}_t$) no se ve distorsionada por las realizaciones extremas de la perturbación sanitaria.

5.3. Metodología de validación fuera de la muestra

Para medir la precisión fuera de la muestra se evaluará en pseudo-tiempo real –simulando que no se conocen los datos posteriores al instante t de la estimación– en un conjunto de métricas estadísticas. Dicho esto, se definen las distintas métricas y pruebas estadísticas que se van a utilizar en este capítulo para ordenar las distintas especificaciones del modelo obtenidas por la prueba del selector paralelizado. La extracción de las predicciones se ha ejecutado mediante una ventana recursiva expansiva, que se define como lo siguiente, partiendo de un instante t inicial, el Filtro de Kalman reestima secuencialmente los parámetros del espacio de estados y proyecta el valor del PIB añadiendo un nuevo periodo a la muestra en cada iteración. Este método replica el proceso real de acumulación de información en

la práctica garantizando que ninguna predicción futura contamine la extracción de la señal presente. Además, se simula el retraso de publicación real que se tendría en el día 30 después de la finalización del trimestre, escondiendo los valores de las variables mensuales cuyo dato se publica más tarde a la llegada de la primera estimación del PIB, en t+30 días desde la finalización del trimestre de referencia.

5.3.1. Ventanas temporales de evaluación

El diseño de la evaluación extramuestral requiere separar la capacidad predictiva estructural del modelo de las distorsiones estadísticas generadas por la pandemia. Por ello, la validación no se realiza sobre una única muestra continua, sino que se divide en dos periodos de evaluación independientes, aislando la perturbación anómala del periodo 2020-2023. Se han delimitado las siguientes ventanas temporales:

- **Ventana pre-COVID (2018-2019, 8 trimestres):** Actúa como escenario de control. Al representar una etapa de estabilidad macroeconómica y baja volatilidad, permite cuantificar el error predictivo base del modelo en condiciones de estado estacionario. Su función es verificar que el Filtro de Kalman no sufre de sobreajuste y opera correctamente en escenarios sin alteraciones estructurales.
- **Ventana pos-COVID (2024-2025, 8 trimestres):** Evalúa el rendimiento del modelo bajo la estructura de covarianzas actual de la economía. El inicio en 2024 se justifica para garantizar que tanto el impacto inicial del confinamiento como el rebote mecánico y la crisis inflacionaria posterior han sido absorbidos por las series.

5.3.2. Métricas de precisión y test

Para evaluar empíricamente la capacidad predictiva del modelo fuera de la muestra, se computan las métricas detalladas en el marco teórico (véase la Sección 3.3.6.3), concretamente la Raíz del Error Cuadrático Medio (RMSE) y el Error Absoluto Medio (MAE).

Para evaluar empíricamente la capacidad predictiva del modelo fuera de la muestra, se computan la Raíz del Error Cuadrático Medio (RMSE) y el Error Absoluto Medio (MAE), siendo el RMSE la métrica principal de selección al penalizar cuadráticamente las grandes desviaciones. Para contextualizar estos resultados, se calculan las ratios relativas frente a dos benchmarks:

- **Modelo ingenuo AR(1):** Un modelo univariante autorregresivo de primer orden que actúa como línea base estadística. Su función es aislar y demostrar el valor añadido real que aporta la inyección de datos de alta frecuencia frente a la simple inercia histórica del PIB.
- **Modelo de la AIReF:** Las proyecciones oficiales del Modelo Factorial Dinámico de la Autoridad Independiente de Responsabilidad Fiscal (MIPred), publicadas desde el año 2017. Representan el estándar institucional más alto a nivel nacional.

Un RMSE relativo inferior a la unidad indicará una ganancia neta de eficiencia predictiva del modelo propuesto frente a estos estándares de referencia.

Además de minimizar el error de predicción, la utilidad de un modelo de *nowcasting* reside en su capacidad para anticipar los puntos de giro del ciclo económico (fases de aceleración o contracción). Para medir esta propiedad se emplea el **índice de concordancia (IC)**, que evalúa la tasa de acierto direccional del signo entre la estimación y la realización observada. Si se define la función indicadora $I(\cdot)$, que toma el valor 1 si la condición se cumple y 0 en caso contrario, el estadístico se formula como:

$$IC = \frac{1}{H} \sum_{h=1}^{H} I\left(\text{sgn}(y_{t+h}) = \text{sgn}(\hat{y}_{t+h|t})\right)$$

Un valor de $IC = 1{,}000$ implica que el modelo ha capturado correctamente la dirección de todas las variaciones trimestrales del periodo analizado. Finalmente, la insesgadez y la eficiencia de las proyecciones se contrastan mediante la ecuación propuesta por Mincer y Zarnowitz (1969). A diferencia de las métricas anteriores, evaluadas sobre ventanas de 8 trimestres, este contraste exige un mayor tamaño muestral para dotar de suficientes grados de libertad a la estimación por Mínimos Cuadrados Ordinarios (MCO) y garantizar la potencia estadística del test. Por consiguiente, esta evaluación específica se ejecuta sobre ventanas temporales ampliadas (2015-2019 para el periodo pre-pandémico y 2021-2023 para el régimen posterior):

$$y_{t+h} = \alpha + \beta \hat{y}_{t+h|t} + u_{t+h}$$

Bajo el supuesto de optimalidad predictiva, se somete a contraste la hipótesis nula conjunta $H_0 : \alpha = 0, \beta = 1$. El no rechazo de esta hipótesis confirma que las proyecciones carecen de sesgo sistemático ($\alpha = 0$) y que el modelo extrae y utiliza la información disponible de forma eficiente ($\beta = 1$).

5.4. Selección del panel de indicadores y optimización predictiva

Una vez delimitadas las ventanas de intervención y las estrategias para mitigar la volatilidad extrema inducida por la crisis de la COVID-19, el objetivo se desplaza hacia la buscar el conjunto de variables con mayor poder predictivo, no meter ruido al modelo. Para resolver esta restricción se diseñó un algoritmo paralelizado de criba exhaustiva. Este selector evalúa de manera recursiva la capacidad predictiva fuera de la muestra en pseudo-tiempo real, simulando el flujo de información y los decalajes institucionales reales a los que se enfrenta un analista en el día a día de la entidad. El diseño del algoritmo estructuró la búsqueda a través de tres fases consecutivas:

1. **Definición de los núcleos fijos de control:** Para garantizar que el espacio latente represente fielmente el ciclo real de la economía española, y con base en el mapa de correlaciones cruzadas, se diseñaron tres núcleos fijos alternativos de 7 variables sectoriales clave (además del PIB trimestral) que el algoritmo estaba obligado a mantener en todas sus iteraciones:
   - **Grupo A (Núcleo de Economía Real - Hard Data):** Compuesto por las variables de empleo efectivo (`AFIL_SS` y `OCU`), producción física (`IPI` e `ICN_SERV`) y flujos de demanda agregada (`VGE_TOT, IMP_BIENES` y `EXP_BIENES`).
   - **Grupo B (Núcleo de Consumo y Expectativas - Soft Data):** Enfocado en indicadores de sentimiento y opinión (`ISE` e `ICC`), variables de comercio minorista (`ICM` y `MAT_TUR`) y servicios de alta frecuencia (`PMI_SERV` y `PASAJ_AER`).
   - **Grupo C (Núcleo Industrial y de Infraestructuras):** Centrado en el sector constructivo e inmobiliario (`VIS, IPI_CONS` y `ICC_CONS`), demanda energética (`CONS_ELEC`) y logística pesada (`MAT_CARGA`).
2. **Generación del espacio combinatorio de búsqueda:** Para cada uno de los tres núcleos de control, las 24 variables restantes de la base de datos original se sometieron a un proceso combinatorio exhaustivo agrupadas de tres en tres. El bucle iterativo generó un espacio de búsqueda compuesto por 2,024 subpaneles únicos candidatos por cada grupo competitivo[4], evaluando la

[4]El número de combinaciones de las variables candidatas se obtiene mediante el coeficiente binomial de la sección transversal: $C_{24,3} = \frac{24!}{3!(24-3)!} = 2{,}024$. Sin embargo, la dimensión real del cómputo es mayor. Para cada una de las 6,072 especificaciones base ($2{,}024 \times 3$ núcleos), el modelo debió reestimarse recursivamente desde cero a lo largo de 16 trimestres de evaluación fuera de la muestra (8 de la ventana pre-COVID y 8 de la ventana pos-COVID) para simular el entorno de pseudo-tiempo real. Multiplicando este proceso secuencial por las 5 estrategias de tratamiento de la pandemia evaluadas (panel bruto con COVID, inserción de valores ausentes en el periodo-1-covid, inserción de valores ausentes en el periodo-2-covid, imputación ARIMA en el periodo-1-covid e imputación ARIMA en el periodo-2-covid), el selector paralelizado ejecutó un total de 485,760 ($6{,}072 \times 16 \times 5$) optimizaciones numéricas completas del algoritmo EM en el espacio de estados.

aportación marginal de información de cada conjunto de indicadores.

3. **Función de jerarquización y ordenamiento:** El algoritmo evaluó el rendimiento de cada subpanel combinatorio bajo las distintas estrategias de control de volatilidad de la pandemia. La regla de selección para ordenar el conjunto de modelos candidatos no se basó en el ajuste histórico o intramuestral, sino en la **minimización de la Raíz del Error Cuadrático Medio ($RMSE$)** fuera de la muestra, también se han tenido en cuenta las otras métricas y la estabilidad ys sentido económico de las cargas factoriales para descartar un modelo con una capacidad predictiva superior a otro pero no cumplía el mínimo necesario en cuanto a estabilidad paramétrica.

5.4.1. selección del modelo final

Tras completar las iteraciones del bucle paralelizado, los modelos se clasificaron según su precisión predictiva en las dos ventanas de control extramuestral independientes: la ventana pre-COVID (2018–2019) de baja volatilidad y la ventana pos-COVID (2024–2025) adaptada al régimen macroeconómico actual. Los resultados de las cinco especificaciones finalistas del "Top 5"se sintetizan la Tabla 5.2.

Tabla 5.2: Análisis comparativo de capacidad predictiva fuera de la muestra (Top 5 finalistas)

.

	Ventana Pre-COVID (2018–2019)					**Ventana Pos-COVID (2024–2025)**				
Modelo	**RMSE**	**vs AR(1)**	**vs AIReF**	**MAE**	**IC**	**RMSE**	**vs AR(1)**	**vs AIReF**	**MAE**	**IC**
M04	**0.185**	**0.978**	**0.774**	**0.145**	**1.000**	**0.155**	**0.411**	**0.630**	**0.120**	**1.000**
M02	0.214	1.132	0.895	0.165	1.000	0.163	0.432	0.662	0.102	1.000
M05	0.245	1.025	1.025	0.191	1.000	0.165	0.437	0.670	0.127	1.000
M01	0.187	0.989	0.782	0.126	1.000	0.183	0.485	0.743	0.142	1.000
M03	0.212	1.121	0.887	0.172	1.000	0.242	0.641	0.983	0.194	1.000

Nota de especificación y variables añadidas al núcleo fijo del Grupo A: El modelo **M04** implementa la ventana de tratamiento del *período-1-covid* mediante proyecciones ARIMA sobre las variables añadidas: `CEM, VGE_SAL, PMI_MANU`, el **M02** aplica valores ausentes sobre el *período-1-covid* con las variables `ISE, VGE_SAL, PMI_MANU` y el **M05** extiende la ventana ARIMA al *período-2-covid* con `PARO, VGE_SAL, PMI_MANU`. El **M01** representa la línea base en datos sin tratamiento sobre las variables `PASAJ_AER, GAS_TUR, PERNOCT_TOT`. El **M03** aplica valores ausentes sobre el *período-2-covid* con `CEM, TURIST_INT, PMI_MANU`. El Índice de Concordancia se denota como IC.

A tenor de las métricas de precisión expuestas, **el Grupo A se consolidó de forma sistemática como el núcleo estructural ganador**. Al cruzar este bloque rígido con las variables añadidas por criba, el modelo **M04** registró el menor error cuadrático de la muestra completa, alcanzando un $RMSE$ mínimo de 0,155 y un MAE medio de 0,120 en la ventana Pos-COVID, batiendo con holgura a la especificación en datos crudos sin tratamiento (M01, $RMSE$ de 0,183).

La ganancia de eficiencia relativa del modelo preferido M04 demuestra una robustez notable frente a los dos exigentes indicadores de comparación: frente a la inercia histórica del modelo univariante ingenuo $AR(1)$, el M04 reduce el error de previsión en un 3 % en el período previo y hasta un 35 % en el régimen Pos-COVID; de igual forma, el modelo propuesto consigue batir las estimaciones en tiempo real publicadas de forma oficial por el modelo MIPred de la AIReF en ambas ventanas temporales. Asimismo, el estadístico de acierto direccional ratifica la consistencia cíclica de la señal, obteniéndose un Índice de Concordancia perfecto ($IC = 1{,}000$) en la totalidad de las especificaciones finalistas, lo que garantiza que el factor latente anticipa con absoluta precisión el signo de los puntos de giro macroeconómicos.

Consecuentemente, el panel de indicadores de la entidad se reduce de las 32 variables iniciales a un subconjunto homogéneo y parsimonioso de 11 variables explicativas. La Tabla 5.3 resume la composición definitiva del modelo de frecuencias mixtas seleccionado.

Tabla 5.3: Panel final de variables (Modelo M04).

.

Sector Económico	**Indicador Macroeconómico (Código)**	**Frecuencia**	**Rol en el Algoritmo**
Variable Objetivo	Producto Interior Bruto (`PIB`)	Trimestral	Núcleo Estructural
Mercado Laboral	Afiliaciones a la Seguridad Social (`AFIL_SS`)	Mensual	Bloque Fijo Base
	Ocupados EPA (`OCU`)	Trimestral	Bloque Fijo Base
Industria	Índice de Producción Industrial (`IPI`)	Mensual	Bloque Fijo Base
	Consumo de Cemento (`CEM`)	Mensual	Añadido por Criba
	Índice PMI Manufacturero (`PMI_MANU`)	Mensual	Añadido por Criba
Consumo	Ventas de Grandes Empresas Totales (`VGE_TOT`)	Mensual	Bloque Fijo Base
	Ventas de G.E. sin Asalariados (`VGE_SAL`)	Mensual	Añadido por Criba
Sector Exterior	Importaciones de Bienes (`IMP_BIENES`)	Mensual	Bloque Fijo Base
	Exportaciones de Bienes (`EXP_BIENES`)	Mensual	Bloque Fijo Base
Sector Servicios	Índice de Cifra de Negocios (`ICN_SERV`)	Mensual	Bloque Fijo Base

Seguidamente, una vez determinada la selección final para el tratamiento de la COVID, resulta necesario validar las perturbaciones antes de proceder a la estimación conjunta del espacio de estados. Siguiendo la metodología de Box-Jenkins para la diagnosis residual (véase la Sección A.5.3 del Apéndice A), se han sometido los errores de predicción de los modelos ARIMA calculados para el panel definitivo a una batería de contrastes de diagnóstico, cuyos resultados se pueden ver en la Tabla 5.4.

Tabla 5.4: Diagnosis paramétrica univariante individual sobre los residuos de imputation (Modelo M04).

.

Variable	**Ljung-Box (Q_p)**	**Media Nula (t_p)**	**Test estacional (QS_p)**	**Shapiro-Wilk (W_p)**	**Signif. Coef (z_p)**
`AFIL_SS`	10.5692 (0.5661)	-0.5944 (0.5529)	0.0000 (1.0000)	0.9621 (0.0000)	¡0.0500
`VGE_TOT`	4.7776 (0.9650)	-0.1143 (0.9091)	0.0000 (1.0000)	0.9370 (0.0000)	¡0.0500
`IPI`	3.7198 (0.9880)	0.0332 (0.9735)	0.0000 (1.0000)	0.9760 (0.0001)	¡0.0500
`IMP_BIENES`	18.5386 (0.1003)	0.0095 (0.9924)	0.0000 (1.0000)	0.9879 (0.0127)	¡0.0500
`EXP_BIENES`	11.2601 (0.5068)	-0.0106 (0.9915)	0.0000 (1.0000)	0.9869 (0.0075)	¡0.0500
`ICN_SERV`	3.8753 (0.9856)	1.1904 (0.2350)	0.8526 (1.0000)	0.8934 (0.0000)	¡0.0500
`CEM`	17.9638 (0.1168)	0.0850 (0.9323)	0.0000 (1.0000)	0.9538 (0.0000)	¡0.0500
`VGE_SAL`	4.1475 (0.9806)	0.0718 (0.9428)	0.8785 (1.0000)	0.8179 (0.0000)	¡0.0500
`PMI_MANU`	12.3330 (0.4193)	-0.3288 (0.7426)	0.0000 (1.0000)	0.9865 (0.0221)	¡0.0500
`PIB`	0.3212 (0.9884)	-0.8357 (0.4053)	0.0000 (1.0000)	0.4591 (0.0000)	¡0.0500
`OCUPADOS`	1.0790 (0.8976)	-0.4078 (0.6849)	0.0000 (1.0000)	0.9653 (0.0810)	¡0.0500

Nota: Valores fuera de los paréntesis indican el estadístico muestral de referencia; los valores entre paréntesis corresponden a sus respectivos p-valores alternativos (p). Nivel de seguridad de referencia: $\alpha = 0{,}05$. Elaboración propia a partir de las estimaciones del algoritmo en R.

Como se ven los resultados en la Tabla 5.4, las perturbaciones univariantes estimadas para anclar el panel refrendan un comportamiento satisfactorio. En lo que respecta a la inercia temporal, la totalidad de los indicadores mensuales y trimestrales superan con holgura el umbral crítico del test de Ljung-Box ($p > 0{,}05$), indicando una ausencia de autocorrelación serial conjunta. De igual forma, el contraste de media nula confirma la ausencia de sesgos sistemáticos en los residuos, concentrándose la totalidad de las probabilidades en la zona de aceptación de la hipótesis nula ($\mathbb{E}[v_t] = 0$). Asimismo, los p-valores del test QS confirman que el preajuste ha limpiado completamente la componente estacional de las

series. Por su parte, el contraste de normalidad de Shapiro-Wilk muestra resultados diferentes. Si bien indicadores de expectativas como el `PMI_MANU` (p = 0,0221) o OCU (`OCUPADOS`, p = 0,0810) parecen tener un perfil cercano al gaussiano, los indicadores contables de retribución y el propio Producto Interior Bruto (`PIB`, W = 0,4591) rechazan de forma categórica la hipótesis de normalidad. Este comportamiento no invalida las proyecciones univariantes, sino que responde a la asimetría y el exceso de curtosis de los indicadores económicos en las colas de la distribución.

5.4.2. Contraste de optimalidad de Mincer-Zarnowitz

La viabilidad de las distintas especificaciones del modelo se sometieron a la prueba de regresión de Mincer y Zarnowitz (1969) anteriormente definida. Cabe destacar que a diferencia de las métricas de error calculadas sobre ventanas temporales de 8 trimestres, este contraste exige un volumen significativamente mayor de grados de libertad. Una muestra excesivamente reducida invalidaría las propiedades estadísticas de los Mínimos Cuadrados Ordinarios (MCO), provocando que los estimadores pierdan consistencia, inflen sus errores estándar y carezcan de potencia estadística para ofrecer un resultado interpretable. Por ello, la regresión se ha ejecutado sobre ventanas temporales ampliadas (2015–2019 para el periodo previo y 2021–2023 para el periodo pos-COVID), garantizando así la regularidad estadística y la robustez del contraste.

Los resultados constituyen una prueba sobre el impacto de la pandemia en la estimación del estado latente. Al observar el modelo base que incorpora la serie histórica sin tratamiento (**M01**), los contrastes en la ventana Pos-COVID ampliada rechazan la hipótesis nula conjunta: el modelo exhibe un sesgo sistemático (p-valor$_\alpha$ = 0.013) y una severa ineficiencia predictiva (p-valor$_\beta$ = 0.013). Este hallazgo demuestra que la absorción de la varianza extrema de 2020 contamina las matrices de covarianza.

Por el contrario, la evaluación de la especificación seleccionada (**M04**) arroja resultados estadísticamente positivos. En la ventana temporal Pos-COVID, el parámetro de sesgo converge a cero (α = 0.002, p-valor = 0.290), confirmando la insesgadez de la predicción al no existir evidencia empírica para rechazar la hipótesis nula. Simultáneamente, el otro parámetro converge prácticamente a la unidad (β= 1.007, p-valor = 0.977), lo que impide rechazar la hipótesis de eficiencia en el uso de la información. Como corolario, el modelo M04 no solo demuestra ser insesgado y eficiente bajo una muestra más grande, sino que logra explicar un $62,4\,\%$ ($R^2_{\text{adj}} = 0.624$) de la varianza del crecimiento intertrimestral fuera de la muestra del periodo Pos-Covid, un valor interesante si lo comparamos con el obtenido por modelos como el del Banco de España *Spain-STING* del año 2021, antes de introducir la volatilidad estocástica en 2024, el cual presentaba un R^2_{adj}=0.73, superando a este modelo pero presentando buenas métricas.

5.5. Selección de parámetros estructurales del modelo

Una vez seleccionado el panel final de 11 variables, se procede a la determinación del número de factores latentes (r) y el orden de retardos autorregresivos (p) que rigen la ecuación de transición.

5.5.1. Determinación del número de factores latentes r

Para la estimación del número de factores latentes se emplean varios de los criterios presentados en la Sección 3.3.4. En primer lugar, se han utilizado los criterios de información asintóticos de Bai y Ng (2002), sin embargo, estos criterios tienden a sobrestimar el número de factores en caso de que el número de variables y de observaciones sea relativamente pequeño, como en este caso, con 11 variables y 376 meses de historia. Debido a esto, se ha utilizado el criterio visual a través del *screeplot* y la Ratio de Autovalores de Ahn y Horenstein (2013)[5].

La Figura 5.2 revela la presencia de un factor fuertemente dominante. El primer componente principal explica, por sí solo, más del 33 % de la varianza total conjunta de las 11 variables. A partir de

[5]Este criterio estadístico determina el número óptimo de factores latentes (r) identificando el punto exacto donde se maximiza el cociente entre autovalores adyacentes (λ_i/λ_{i+1}) de la matriz de covarianzas.

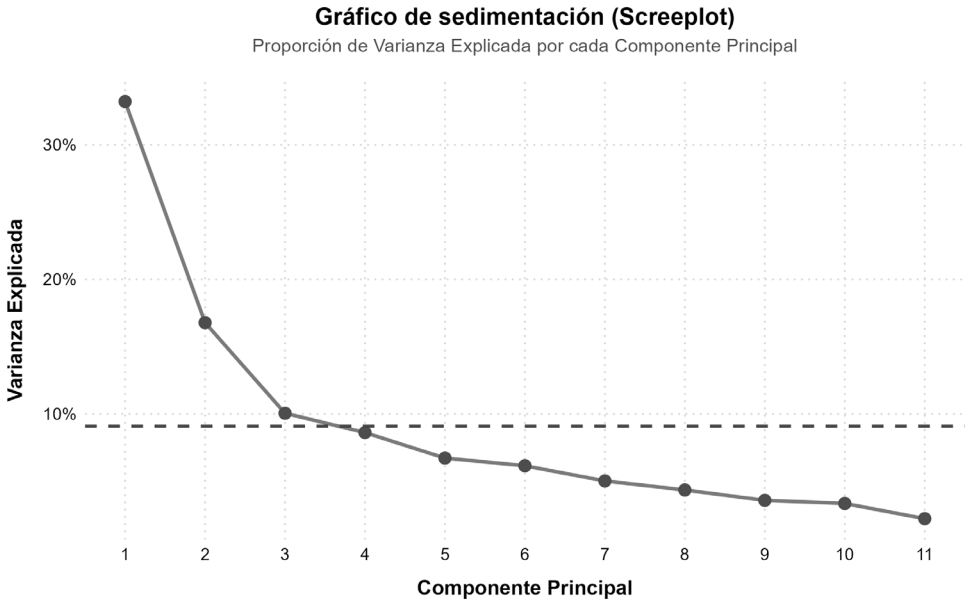


Figura 5.2: Gráfico de sedimentación: Proporción de Varianza Explicada por componente

.

este punto, se observa una caída abrupta (el conocido "codo") hacia el segundo componente ($\approx 16{,}8\,\%$). La línea horizontal discontinua marca el umbral de información media individual ($1/N \approx 9{,}09\,\%$); los componentes que subyacen a este nivel se consideran ruido.

Además de la evidencia visual, también se utilizó un contraste a través del estadístico de la Ratio de Autovalores. Como se detalla en la Tabla 5.5, la evaluación de los componentes principales extraídos confirma la especificación de un único factor, r=1. El estadístico alcanza su máximo global en el primer componente ($r = 1$, Ratio = 1,99), superando al segundo (Ratio = 1,70) y evidenciando una estabilización en los vectores subsiguientes.

Tabla 5.5: Contraste de Ratio de Autovalores

.

Componente (i)	**Autovalor (λ_i)**	**Ratio de Autovalores (ER_i)**
1	**3,68**	**1,99**
2	1,85	1,70
3	1,08	1,20
4	0,90	1,17
5	0,77	1,21

Nota: El estadístico ER_i se calcula como el cociente λ_i/λ_{i+1}. El número óptimo de factores se determina maximizando la secuencia de dicho estadístico. Elaboración propia a partir de los resultados del algoritmo.

En consecuencia, el valor del estadístico ER lleva a la selección de un único factor dinámico ($r = 1$)

que permite capturar el comovimiento estructural del panel como un único pulso de la economía.

5.5.2. Selección del orden autorregresivo de la ecuación de transición (p)

Para determinar el orden de retardos (p) del factor común latente ($r = 1$) en la ecuación de transición, se analizó la trayectoria del factor estimado mediante criterios de información formales. Se aplicó la función de optimización multivariante `VARselect` (paquete `vars` (Pfaff, 2008)) sobre la serie del factor obtenida por Máxima Verosimilitud. Todos los criterios de información asintóticos —AIC, HQ, SC/BIC y FPE[6]— convergieron de manera unánime en la selección de un orden óptimo $p = 2$.

De este modo, la dinámica del factor común quedó parametrizada bajo la siguiente especificación de transición:

$$F_t = 0{,}8398F_{t-1} - 0{,}0571F_{t-2} + u_t. \tag{5.1}$$

Siguiendo el manual de (Aneiros, 2025), y con el fin de verificar la estabilidad y reversión a la media del sistema, se evaluó el polinomio característico asociado a la ecuación (5.1), cuya definición teórica general se detalla en el Apéndice A.4 (Definición A.15):

$$z^2 - 0{,}8398z + 0{,}0571 = 0.$$

Aplicando la fórmula general de resolución cuadrática sobre los coeficientes estimados del factor ($a = 1$, $b = -0{,}8398$ y $c = 0{,}0571$), se obtiene:

$$z = \frac{0{,}8398 \pm \sqrt{(-0{,}8398)^2 - 4(1)(0{,}0571)}}{2} = \frac{0{,}8398 \pm 0{,}690553}{2}.$$

Dando como resultado las dos raíces características del sistema dinámico:

$$z_1 \approx 0{,}7652, \quad z_2 \approx 0{,}0746.$$

La validez y robustez de esta especificación se justifica mediante dos propiedades algebraicas clave:

- **Estabilidad asintótica ($|z_i| < 1$):** Al situarse ambas raíces estrictamente dentro del círculo unitario, se garantiza la estacionariedad del factor. En la representación alternativa de operadores de retardo (Definición A.16, Apéndice A.4), las raíces recíprocas resultan ser $w_1 \approx 1{,}307$ y $w_2 \approx 13{,}40$, cumpliendo simétricamente con la condición de causalidad al ubicarse fuera del círculo unitario ($|w_i| > 1$).
- **Con ($D > 0$):** Dado que el discriminante es estrictamente positivo ($D \approx 0{,}4769$), las raíces son reales puras (no complejas conjugadas). Esto asegura que ante cualquier shock exógeno, el factor retorne a su media de forma monótona y suave, evitando fluctuaciones importantes ante la presencia de datos faltantes en el borde dentado del panel.

Esta estructura AR(2) cuenta con un sólido respaldo en la teoría clásica de los ciclos económicos (Hamilton, 1994; Stock y Watson, 2002b), al ser la representación paramétrica mínima capaz de capturar la inercia de las transiciones entre expansiones y recesiones. Asimismo, la especificación se alinea con la práctica de MFD de referencia como el *MIPred* de la AIReF (AIReF, 2024), donde se restringe ex-ante la dinámica del factor a un orden dos para asegurar la convergencia del Filtro de Kalman y que las estimaciones del PIB en tiempo real sean robustas.

[6]Estos indicadores penalizan la pérdida de grados de libertad para evitar el sobreajuste. El AIC y el FPE aplican penalizaciones moderadas; el SC/BIC penaliza de forma severa en función de $\ln(T)$ priorizando la parsimonia; y el HQ establece un término medio.

Capítulo 6

Ajuste, Diagnosis y Análisis Dinámico en Tiempo Real

Una vez determinada la especificación estructural del modelo definitivo (M04), parametrizado bajo un panel de 11 variables unificado con un factor común ($r = 1$) de inercia autorregresiva de segundo orden ($p = 2$), en el presente capítulo se aborda el ajuste intramuestral global y la evaluación con ventana recursiva en tiempo real. Un modelo de reducción de la dimensión solo demuestra su validez si es capaz de generalizar su aprendizaje histórico e interpretar las revisiones dinámicas conforme la información fluye a al realidad.

Este capítulo aborda el ajuste intramuestral y la evaluación predictiva en tiempo real del modelo M04. A lo largo de las siguientes secciones se analiza la sincronización del factor latente frente al PIB, su capacidad explicativa y su estabilidad ante quiebres estructurales. Posteriormente, se evalúa el error predictivo intra-trimestral mediante la descomposición algebraica de noticias de Bańbura y Modugno (2014) y se somete el modelo a pruebas de estrés.

6.1. Ajuste e Inferencia del Modelo Factorial Dinámico

Una vez determinada la representación estructural del modelo, con un factor único ($r = 1$) con inercia de segundo orden ($p = 2$), se procede a la estimación definitiva de los parámetros del modelo con un tamaño muestral que va desde febrero de 1995 hasta abril de 2026, último dato obtenido cerrando el periodo de obtención de datos el 21 de Mayo de 2026.

El ajuste se ejecuta mediante la función `DFM()` del paquete `dfms` en el entorno R. En cuanto al método de estimación, se ha configurado el método EM de *Bańbura-Modugno* (`em.method = "BM"`), optimizado para lidiar con el patrón de datos faltantes y las diferencias de publicación. No se ha restringido por valores faltantes (`max.missing = 1`), esto quiere decir que se permite hasta un 100 % de valores faltantes en la fila o lo que es lo mismo que no haya ningún dato del panel, y también se han declarado explícitamente el PIB y los Ocupados como variables de baja frecuencia (`quarterly.vars`), activando la agregación temporal de Mariano y Murasawa (2003) dentro del filtro.

Tras obtener las estimaciones del modelo, la estructura matemática del mismo se presenta en el sistema de ecuaciones en el Espacio de Estados que el lector puede ver con mayor detenimiento en el Apéndice C

6.1.1. Análisis del Ajuste Intramuestral del Factor Común

Para evaluar la capacidad del modelo a la hora de capturar las fluctuaciones reales de la actividad productiva, se analiza el comportamiento del factor común estimado frente a la variable de interés, el PIB. El factor dinámico inobservable se extrae originalmente con frecuencia mensual, media cero

y varianza unitaria. Por lo tanto, se necesita transformar para ser comparable al PIB en variaciones trimestrales.

En primer lugar, se aplica la agregación de Mariano y Murasawa (2003), lo que permite transformar la señal mensual del factor a frecuencia trimestral:

$$F_t^Q = \frac{1}{3}F_t + \frac{2}{3}F_{t-1} + F_{t-2} + \frac{2}{3}F_{t-3} + \frac{1}{3}F_{t-4}$$

En segundo lugar, se ejecuta un reescalado para que ambas variables sean comparables. Los parámetros de tipificación se calculan sobre la submuestra histórica pre-pandemia (febrero de 1995 a diciembre de 2019). Este procedimiento impide que la volatilidad extrema del año 2020 distorsione la desviación típica de la serie completa, preservando la escala las fluctuaciones cíclicas en los periodos de normalidad:

$$F_{t,\text{escalado}}^Q = \left(\frac{F_t^Q - \hat{\mu}_{F,\text{pre}}}{\hat{\sigma}_{F,\text{pre}}} \right) \hat{\sigma}_{\text{PIB,pre}} + \hat{\mu}_{\text{PIB,pre}}$$

donde $\hat{\mu}$ y $\hat{\sigma}$ representan las medias y desviaciones típicas de las respectivas series en el periodo de estabilidad pre-pandemia.

La evolución conjunta de ambas series se presenta en la Figura 6.1. Debido a la magnitud histórica del shock de la COVID-19 en el segundo trimestre de 2020 (donde el PIB cayó un -17.79 %), la escala vertical del panel global (Figura 6.1a) se comprime y no se terminan de ver los detalles más interesantes de la comparativa. Para evaluar el ajuste sin esta limitación, Figura de la muestra total (1995-2026) se complementa con dos ventanas de observación: un periodo histórico pre-covid (Figura 6.1b) y un seguimiento del ciclo de normalización reciente, de 2021 a 2026(Figura 6.1c).

A través del análisis visual pueden observar que en los periodos de estabilidad (Figura 6.1b), la trayectoria del factor dinámico y el PIB observado es muy similar, replicando la tendencia media de crecimiento española (en el entorno del 0.8 % - 1.0 % intertrimestral) y anticipando los puntos de giro del ciclo económico. Esta precisión se confirma durante la Gran Recesión de 2008–2009, donde el factor acompaña la caída de la actividad hasta su "valle" en el primer trimestre de 2009 (-2.67 %).

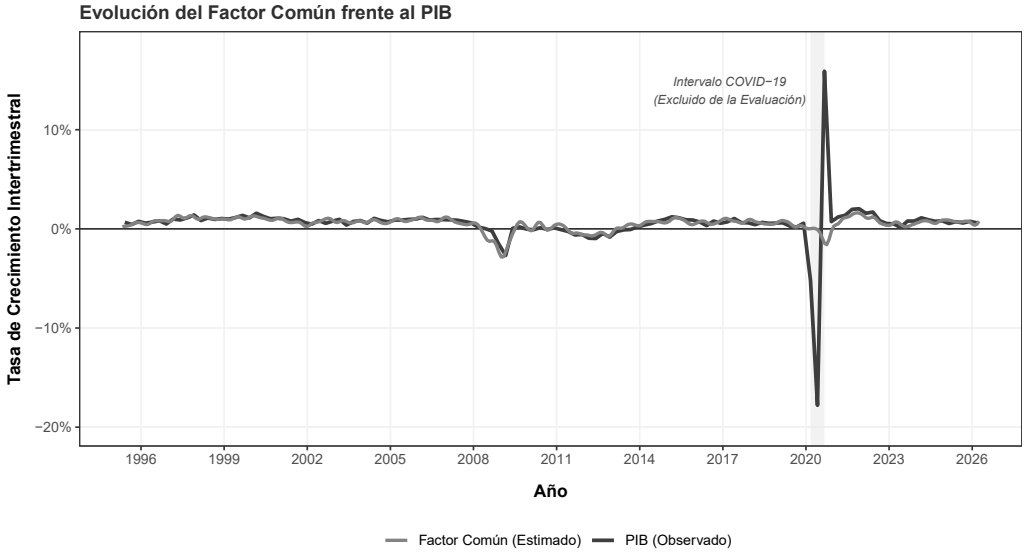
Por su parte, el ciclo post-covid (Figura 6.1c) confirma la estabilidad de la estimación una vez superado el momento de mayor varianza de la serie, identificada en la Sección 5.2. Tras capturar la fase de recuperación de 2021 y la desaceleración producto de la crisis inflacionaria de finales de 2022, el factor común se estabiliza y se sincroniza con el avance real de la economía en la ventana actual (2024–2025), consolidando tasas de crecimiento en el entorno del 0.8 % intertrimestral.

A través del análisis exploratorio que se acaba de presentar ya se puede observar cierta relación en la evolución de ambas series pero para cuantificar el grado de sincronización y analizar la estructura de dependencias temporales entre el factor latente y la actividad productiva, se estimó la función de correlación cruzada (CCF) bajo tres escenarios macroeconómicos bien definidos: la ventana de estabilidad pre-pandemia (1995–2019), la muestra completa que incorpora las distorsiones del año 2020 con el tratamiento de relleno con ARIMA (1995–2026), y la fase post-confinamiento (desde agosto de 2020 hasta 2026). Los coeficientes estimados para los distintos retardos y adelantos trimestrales se presentan de forma comparativa en la Tabla 6.1.

El análisis de las tres submuestras permite extraer diferencias notables entre periodos:

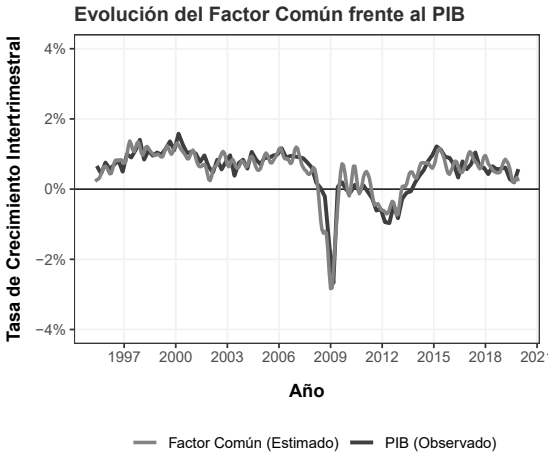
El análisis de las tres submuestras revela varios comportamientos clave en el ajuste del modelo. Durante el periodo de estabilidad pre-pandemia (1995–2019), la correlación contemporánea alcanza un 0,9044, confirmando al factor como un indicador coincidente adecuado. Sin embargo, al incluir las observaciones de la pandemia en la muestra total, esta correlación cae abruptamente hasta el 0,1422. Esta ruptura geométrica se explica por la inflación de la varianza del PIB oficial frente a la estabilización del factor latente intervenido mediante el tratamiento ARIMA. Finalmente, al aislar la ventana post-COVID, el coeficiente se restituye con fuerza (0,8377), lo que demuestra que la perturbación sanitaria no causó daños estructurales permanentes en las propiedades del factor.

Los resultados del análisis realizado confirman esta trayectoria de normalización. Al aislar el periodo post-COVID (septiembre de 2020–2026), la correlación contemporánea se restituye con fuerza,



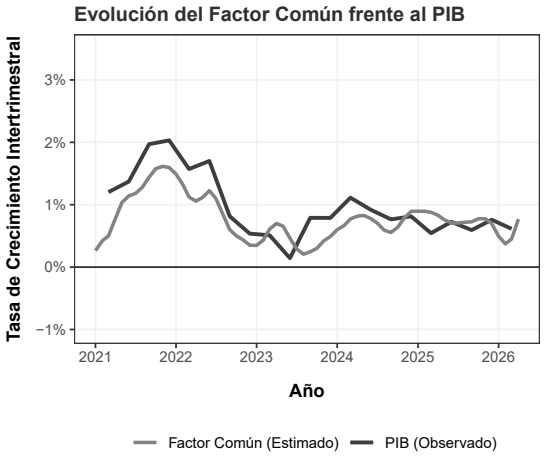
(a) Evolución histórica completa (1995–2026)

.



(b) Ventana histórica ordinaria (1995–2019)

.



(c) Ciclo reciente post-pandemia (2021–2026)

.

Figura 6.1: Ajuste intramuestral del Factor Común frente a las tasas del PIB Real

.

Tabla 6.1: Función de correlación cruzada (CCF) entre el Factor Común y el PIB Real

.

Trimestre (k)	**Pre-COVID (1995–2019)**	**Muestra Total (1995–2026)**	**Post-COVID (2021–2026)**	**Post-COVID Est. (2022–2026)**
-4	0.4827	0.0425	-0.1716	-0.3953
-3	0.6108	0.1379	0.0650	-0.2713
-2	0.7095	0.1667	0.3851	-0.1135
-1	0.8320	0.2184	0.6097	0.0085
0	**0.9044**	**0.1422**	**0.8377**	**0.6413**
$+1$	0.7478	0.3502	0.7168	0.2214
$+2$	0.5523	0.2854	0.4603	-0.0100
$+3$	0.4616	0.2008	0.2273	0.1106
$+4$	0.3894	0.1163	-0.0826	-0.1136

Nota: El retardo $k < 0$ indica que la dinámica del factor mensual precede a la publicación del PIB trimestral. Elaboración propia a partir de los resultados de estimación en R.

alcanzando un valor de 0,8377, mientras que la asimetría temporal del liderazgo del factor vuelve a su comportamiento habitual pre-COVID ($\rho = 0{,}6097$ en $k = -1$ frente a $\rho = 0{,}7168$ en $k = 1$). Al acotar el análisis a la ventana de estabilidad posterior al periodo-2-covid identificado anteriormente(abril de 2022–2026), el coeficiente se asienta en un valor 0,6413, algo inferior al del anterior periodo analizado. Esta rápida convergencia demuestra que la perturbación de la crisis sanitaria no causó daños estructurales permanentes en las propiedades del factor y confirma que, una vez depurado el shock univariante de los meses centrales, la señal inobservable sigue guiando con precisión el comportamiento coyuntural de la economía española actual.

Para verificar formalmente la dirección de esta causalidad se aplicó el test de causalidad de Granger (Granger, 1969)[1]. Los resultados de los contrastes se detallan en la Tabla 6.2.

Tabla 6.2: Resultados del contraste de causalidad de Granger

.

Hipótesis Nula (H_0)	**Estadístico F**	**p-valor**	**Decisión (al 1 %)**
El Factor Común no causa al PIB	22.080	$8{,}822 \times 10^{-6}$	Rechaza H_0 (Sí causa)
El PIB no causa al Factor Común	0.0009	0.9766	No rechaza H_0

Nota: Contraste estimado con un orden de retardo $p = 2$ determinado por criterios de información sobre la submuestra histórica pre-pandemia (1995–2019). Elaboración propia.

Este resultado es de suma importancia para la validez del modelo al confirmar que la relación de causalidad funciona en un único sentido -donde el factor dinámico precede y anticipa al PIB, pero el PIB no influye en la trayectoria del factor-, se valida el diseño matemático del espacio de estados.

[1] El test de causalidad de Granger (Granger, 1969) evalúa si el comportamiento histórico de una variable (X_t), aporta información estadísticamente relevante para predecir la trayectoria de otra (Y_t), más allá de la información ya contenida en la propia inercia autorregresiva de esta última. Formalmente, se estima un modelo dinámico bivariante mediante mínimos cuadrados y se contrasta la hipótesis nula de no causalidad ($H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0$) sobre la ecuación de proyección:

$$Y_t = \alpha_0 + \sum_{i=1}^{p} \alpha_i Y_{t-i} + \sum_{j=1}^{p} \beta_j X_{t-j} + u_t,$$

donde el orden de rezagos p se determina mediante criterios de información (como AIC o BIC). Un rechazo de la hipótesis nula mediante un estadístico F implica que la trayectoria de la variable X_t precede temporalmente y con carácter predictivo a la variable Y_t

Desde el punto de vista metodológico, esto garantiza que la ecuación de transición es verdaderamente independiente. Al descartar cualquier efecto de retroalimentación o causalidad bidireccional, el modelo queda libre de sesgos de endogeneidad, garantizando la consistencia de las estimaciones del factor y de las previsiones del PIB. Por último, se introduce el análisis de dispersión contemporáneo de ambas series en la Figura 6.2, donde se cruza la tasa del factor común frente al PIB real, identificando las observaciones del periodo de confinamiento de la COVID-19 (Q1 a Q4 de 2020) en color rojo.

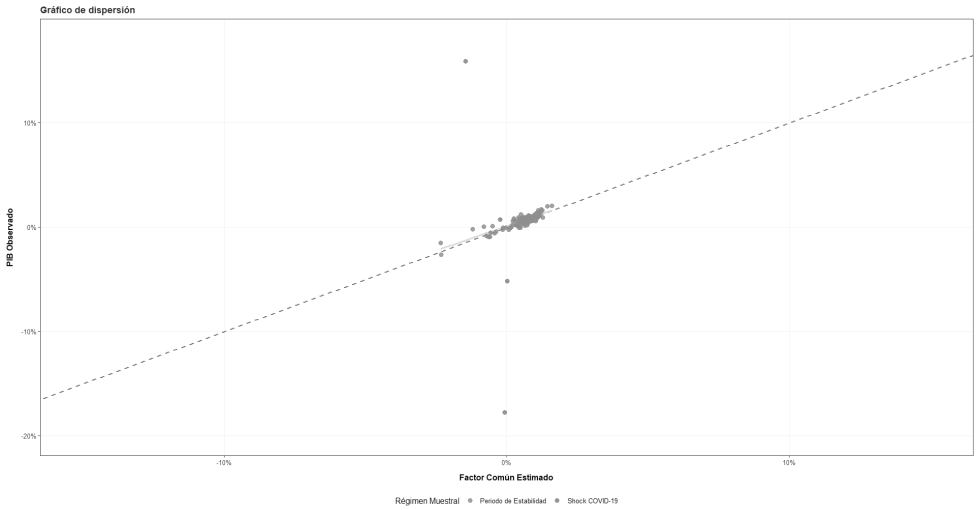


Figura 6.2: Figura de dispersión del Factor Común frente al PIB Real

.

El análisis de la Figura 6.2 permite constatar la estabilidad del modelo. Las observaciones del periodo considerado "normal" o el complementario del periodo COVID (puntos azules) se concentran de manera homogénea a lo largo de la recta de ajuste calculada por Mínimos Cuadrados Ordinarios (MCO) (reflejada en verde en la Figura) y de la diagonal de identidad ($X = Y$), lo que demuestra que el modelo opera de forma insesgada bajo regímenes "normales" del ciclo económico. Por su parte, los registros del *shock* de la COVID-19 (puntos rojos) ilustran el impacto del tratamiento ARIMA. Mientras las observaciones de marzo (Q1) y diciembre (Q4) se sitúan cerca de la normalidad histórica, los trimestres de confinamiento y rebote (Q2 y Q3) muestran desviaciones críticas. Debido a la estabilización univariante aplicada sobre el panel de indicadores entre marzo y julio, el Filtro de Kalman contuvo la varianza del factor común en torno a su media, situándolo cerca del eje central horizontal del diagrama. Consecuentemente, el colapso del PIB oficial hasta el -17.79 % en el segundo trimestre y su posterior repunte del 15.90 % en el tercero se proyectan de forma puramente vertical, completamente alejado de la diagonal de identidad. Esta separación confirma que la estrategia de tratamiento para la COVID-19 impidió que el factor latente cayera de forma acusada ante el primer confinamiento, evitando un efecto palanca estructural que habría distorsionado la estimación de los parámetros para el resto de la serie histórica.

Una vez contrastada esta estabilidad global frente a observaciones atípicas, se analiza la estructura interna de los parámetros estimados que rigen esta relación. En este sentido, el vector de cargas factoriales (Λ) constituye el parámetro principal de la ecuación de observación, al cuantificar la sensibilidad y el grado de co-movimiento de cada indicador individual respecto al ciclo económico común. La Figura 6.3 recoge las cargas estimadas para el panel de variables, ordenadas según su contribución relativa, incorporando una línea discontinua que señala el impacto absoluto medio como umbral de referencia.

Antes de desglosar el peso de los indicadores, es preciso señalar una particularidad algebraica del modelo. A diferencia de otras especificaciones estáticas que restringen la varianza del factor latente a la unidad forzando cargas absolutas elevadas, el algoritmo de Cuasi-Máxima Verosimilitud del paquete

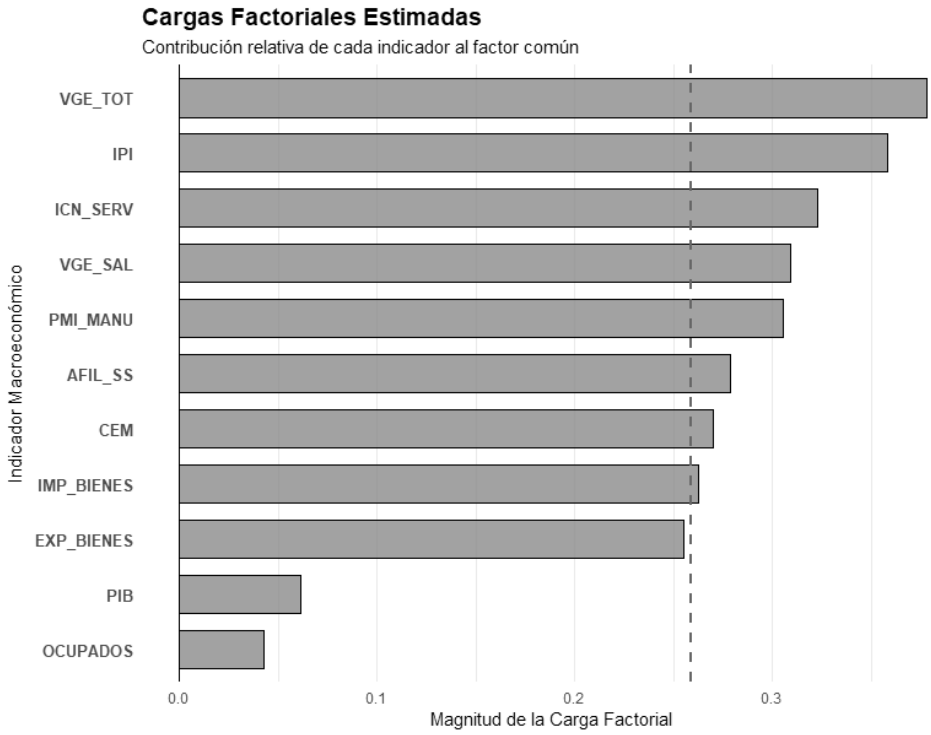


Figura 6.3: Cargas Factoriales de las Covariables en el Modelo

.

`dfms` aplica criterios de normalización matricial sobre las propias cargas. Esto distribuye el peso de forma acotada en un rango de 0,04 a 0,38. Lo relevante para el análisis, por tanto, no es la escala nominal absoluta, sino la aportación relativa y el signo de cada coeficiente. La evaluación de las cargas factoriales muestra las siguientes dinámicas estructurales: las variables vinculadas a la actividad industrial y comercial (como las Ventas de Grandes Empresas o el IPI) actúan como los motores principales del factor común, situándose por encima de la media. Por su parte, los indicadores de empleo y encuestas de opinión conforman un bloque de aportación intermedia y estable. Por último, los coeficientes asociados a las variables trimestrales exhiben magnitudes más bajas en la ecuación mensual debido a su frecuencia, aunque mantienen su sincronización a través de la agregación temporal. De este modo, aunque tengan poco peso inmediato en la ecuación mensual, el PIB y el empleo se mueven de forma totalmente sincronizada al compás del factor que impulsan los indicadores mensuales. Para comprender el alcance explicativo del factor, resulta lógico analizar la proporción de varianza individual que el factor logra capturar de cada indicador del panel, métrica conocida en la literatura factorial como **comunalidad**.

La Tabla 6.3 presenta la comunalidad (R^2) y la varianza del error de observación ($diag(R)$) para cada variable del panel, arrojados por el ajuste del algoritmo.

Tabla 6.3: Proporción de varianza explicada por el Factor (R^2) y varianza del error (R)

.

Variable	**Varianza Explicada por el Factor (R^2)**	**Varianza del Error (**
Indicadores de frecuencia mensual		
Índice PMI Manufacturero (`PMI_MANU`)	0.4690	0.0769
Afiliaciones a la Seguridad Social (`AFIL_SS`)	0.3837	0.7793
Cifra de Negocios Servicios (`ICN_SERV`)	0.3287	0.7509
Ventas Grandes Empresas Totales (`VGE_TOT`)	0.2403	0.6755
Ventas Grandes Empresas Asalariadas (`VGE_SAL`)	0.2398	0.7852
Índice de Producción Industrial (`IPI`)	0.1838	0.6810
Importaciones de Bienes (`IMP_BIENES`)	0.1397	0.8586
Consumo de Cemento (`CEM`)	0.1260	0.8330
Exportaciones de Bienes (`EXP_BIENES`)	0.0642	0.8164
Indicadores de frecuencia trimestral		
Producto Interior Bruto (`PIB`)	**0.9910**	**0.0201**
Ocupados EPA (`OCUPADOS`)	**0.9905**	**0.0464**

Nota: Resultados extraídos de la estimación del modelo base (`summary`). La varianza explicada para las series trimestrales presenta una inflación artificial derivada de la interpolación autorregresiva de valores ausentes. Elaboración propia.

Los resultados de la Tabla 6.3 confirman que el factor latente captura una porción muy sustancial de la varianza en los indicadores líderes mensuales, especialmente en el PMI Manufacturero (explicando casi el 47 %) y el empleo afiliado (38,3 %). Sin embargo, la atención debe dirigirse a los valores registrados por las variables de baja frecuencia: el R^2 de la Contabilidad Nacional y la EPA asciende a niveles superiores al 0,99, acompañados de unas varianzas del error residual virtualmente nulas (0,0201 y 0,0464, respectivamente).

Adicionalmente, para garantizar la consistencia estadística de la estimación, se analiza la estructura de la matriz de covarianzas de los residuos de la ecuación de observación, denotada como $\hat{\Sigma}_e = \text{Cov}(y_t - \Lambda F_t)$. En un MFD bajo el supuesto de ortogonalidad(MFD Exacto, véase la Sección 3.3.2.3), se asume que el factor común absorbe la totalidad de la varianza conjunta, resultando en una matriz de residuos diagonal. No obstante, en paneles macroeconómicos reales, es habitual la presencia de comovimientos locales o perturbaciones de carácter sectorial. La estimación del modelo en este proyecto revela la existencia de correlaciones residuales estadísticamente significativas (marcadas con asterisco en la salida

del algoritmo) concentradas en dos bloques bien definidos. Por un lado, se identifica un *bloque industrial y de comercio exterior*, reflejado en la covarianza residual positiva y significativa entre el Índice de Producción Industrial (`IPI`) y las Ventas de Grandes Empresas (`VGE_TOT`, 0,3877), así como entre el `IPI` y las Exportaciones (`EXP_BIENES`, 0,2641). Por otro lado, se aprecia un *bloque terciario y laboral*, caracterizado por la relación residual entre las Afiliaciones (`AFIL_SS`), la Cifra de Negocios de Servicios (`ICN_SERV`, 0,1335) y los Salarios de Grandes Empresas (`VGE_SAL`, 0,1025). Este incumplimiento del supuesto no invalida la especificación, la presencia de estas dependencias locales sectoriales justifica la adopción del *Modelo de Factores Dinámicos Aproximado*, cuyos fundamentos teóricos de convergencia y control transdimensional se expusieron en la Sección 3.3.2.3. Bajo este marco teórico, se asume que el estimador Cuasi-EM preserva su consistencia y robustez asintótica siempre que las relaciones cruzadas permanezcan acotadas. La verificación formal de las condiciones de Chamberlain y Rothschild (1983) y Stock y Watson (2002b) para la matriz de datos de este modelo, basada en la acotación espectral de su autovalor máximo y la sumabilidad absoluta de sus filas, se detalla de forma explícita en el Apéndice C. Este análisis confirma que el panel opera bajo un régimen de correlación transversal débil, protegiendo la trayectoria del factor frente a perturbaciones locales entre las variables.Finalmente, destacan las covarianzas residuales asociadas a (`PIB`) y la Encuesta de Población Activa (`OCUPADOS`) frente al resto del panel (registrando valores inferiores a 0,01 en prácticamente todos los cruces mensuales). Esto hace pensar que la señal propia de las variables de frecuencia trimestral se encuentra prácticamente libre de contaminación cruzada, garantizando que el proceso de actualización del Filtro de Kalman se realice sobre proyecciones estables.

6.1.2. Métrica de bondad de ajuste intramuestral

Esta sobreestimación del ajuste justifica la necesidad de aplicar métricas de bondad de ajuste alternativas a las proporcionadas por defecto por el paquete de R utilizado para la estimación del modelo. Esta sobrevaloración de la comunalidad de las variables trimestrales sucede cuando el modelo de Espacio de Estados se especifica incorporando una estructura autorregresiva de primer orden en las perturbaciones específicas de cada serie (`idio.ar1 = TRUE`), el coeficiente R^2 (como el 0,9910 reportado en la Tabla 6.3) se sobrevalora. Este comportamiento no constituye un indicador real de la capacidad predictiva del factor, sino un resultado del Suavizador de Kalman. Al estar el PIB mensualizado y presentar valores ausentes (`NA`) en dos de cada tres observaciones, el suavizador interpolará los meses faltantes utilizando la persistencia AR(1) de los errores de autocorrelación del error e_t. Esta interpolación proyectada sobre el PIB latente en los meses que no hay dato es tan precisa que la senda estimada se ciñe a la trayectoria observada, reduciendo la varianza residual a prácticamente cero e invalidando el R^2, comunalidad en el análisis factorial, como métrica de bondad de ajuste. Para corregir esto y aislar la contribución del factor, en este trabajo se computa el coeficiente $R^2_{\textbf{Factor-only}}$ a través de la proyección ΛF_t, dejando fuera la componente de persistencia autorregresiva aleatoria mediante el argumento de control `use.full.state = FALSE`. El **modelo M04 arroja un ajuste del** 65,43 %. Como documentan Stock y Watson (2002b), una vez eliminada la varianza tendencial y las raíces unitarias espurias mediante la diferenciación de las series, capturar la volatilidad del ciclo y el ruido de las series es una tarea compleja. Que un único factor común sea capaz de absorber de forma directa el 65,43 % de las variaciones intertrimestrales del PIB español confirmael panel elegido se mueve al unísono y valida la eficacia de la estrategia de intervención univariante aplicada durante el confinamiento.

6.1.3. Diagnóstico de los residuos

La validación estadística se completa analizando las propiedades temporales de las innovaciones de la ecuación de medida, definidas algebraicamente como $v_{it} = e_{it} - \rho_i e_{i,t-1}$. Al configurar el modelo dinámico incorporando perturbaciones aleatorias autorregresivas (`idio.ar1 = TRUE`), la consistencia teórica del sistema bajo supuestos ideales (MFD Exacto) requeriría que el vector $\mathbf{v}_t$ se comportase como un ruido blanco vectorial incorrelacionado. Con el fin de validar esta condición, al aplicar el

test de Ljung-Box sobre una malla de 12 retardos mensuales, se rechaza la hipótesis nula de ausencia de autocorrelación serial (p-valor $< 0{,}05$) en la mayoría de los indicadores del panel. Si bien en la modelización univariante de series temporales como Box-Jenkings este resultado evidenciaría una especificación errónea, en el marco asintótico de los MFD de alta dimensión este comportamiento está justificado y no compromete la consistencia de la señal del ciclo por tres razones teóricas:

En primer lugar, como ya se indicó en la Sección 3.3.3.3 bajo el método de estimación por Cuasi-Máxima Verosimilitud (QML), Doz et al. (2012) demostraron formalmente que el estimador del factor común obtenido mediante el algoritmo EM y el suavizador de Kalman es asintóticamente consistente cuando la sección transversal y la dimensión temporal tienden a infinito ($N, T \to \infty$), incluso ante la presencia de especificaciones incorrectas en la dinámica de los residuos. La consistencia del factor latente $\hat{F}_t$ descansa en la ley de los grandes números sobre la sección transversal, la cual promedia y diluye el ruido aleatorio local e inercial frente a la señal común dominante. Cabe precisar que, si bien dicho marco asintótico se formula bajo el supuesto de una sección transversal que tiende a infinito, la adopción de un panel de escala moderada ($N = 11$) en este trabajo responde a una estrategia deliberada de regularización. En muestras finitas, las simulaciones de Monte Carlo presentadas por Doz et al. (2012) ya confirman la robustez numérica y convergencia del algoritmo Quasi-EM en entornos de baja dimensionalidad (v.g., $N = 10$). Bajo estas condiciones de dimensionalidad moderada, la flexibilización del supuesto de ortogonalidad adquiere una necesidad: mientras que en paneles masivos las correlaciones propias locales se cancelan de forma pasiva por simple agregación, en un panel acotado estas dependencias transversales débiles persisten. Su modelización explícita mediante el Filtro de Kalman es, por tanto, un requisito indispensable para purgar los comovimientos sectoriales y aislar la señal del factor común de forma insesgada. En segundo lugar, ante desviaciones del supuesto de **ruido blanco en muestras finitas**, el Filtro de Kalman mantiene intacta su validez como el estimador lineal insesgado de mínima varianza (conocido indistintamente como *MVLUE* o *BLUP*, por sus siglas en inglés). Tal como se fundamentó formalmente en la Sección 3.2.1 a partir del Teorema 2 de Kalman (1960), al restringir la optimización a la clase de proyecciones lineales sobre el espacio de observaciones, el suavizador garantiza la extracción de la trayectoria de mínima variabilidad del error, con independencia de la normalidad o la ortogonalidad temporal de las perturbaciones residuales. En tercer lugar, la restricción de la dinámica de los errores a un proceso AR(1) responde a la dimensional del problema. Bajo la metodología de Box-Jenkins (modelos ARIMA univariantes), la persistencia de autocorrelación residual exigiría la estimación de retardos adicionales (estructuras AR(2) o ARMA) hasta obtener un ruido blanco, dado que el coste en grados de libertad es marginal (un parámetro por serie). Sin embargo, trasladar este enfoque univariante a un modelo de espacio de estados de $N = 11$ variables exigiría expandir las matrices del sistema estimando decenas de parámetros adicionales de forma simultánea. En muestras finitas, el error de estimación derivado de esta sobreparametrización elevaría la varianza global del algoritmo de Cuasi-Máxima Verosimilitud, induciendo problemas de sobreajuste que podrían reducir la capacidad de capacidad predictiva del factor común.

Por consiguiente, se justifica la especificación de AR(1), aun con evidencia de leve correlación serial remanente,se defiende el control del equilibrio sesgo-varianza. Esta aproximación metodológica se alinea con la práctica habitual de la literatura aplicada de tiempo real (Bańbura y Modugno, 2014) y con el modelo para la economía española de Cuevas y Quilis (2012), priorizando la estabilidad y precisión del *nowcasting* frente al ajuste intramuestral. No obstante, como señala la literatura sobre *nowcasting* (Bańbura y Modugno, 2014; Giannone et al., 2008), las métricas de bondad de ajuste intramuestral constituyen únicamente un diagnóstico preliminar. El elevado número de parámetros de estos modelos en el espacio de estados puede derivar en problemas de sobreajuste sobre los datos históricos, por ello, la verdadera validación de la utilidad del Factor Común estimado no viene en su capacidad para explicar el pasado, sino en su precisión para anticipar la variable de interés simulando el flujo de información en tiempo real. Esta evaluación basada en el contraste de los errores de predicción de nuevas observaciones, constituye la métrica central para la evaluación del modelo.

6.1.4. Análisis de estabilidad paramétrica ante shocks estructurales

Una de las asunciones teóricas fundamentales de los Modelos de Factores Dinámicos es la hipótesis de invarianza temporal en sus matrices estructurales. Esta propiedad exige que el vector de cargas factoriales ($\mathbf{\Lambda}$) permanezca estable a lo largo del horizonte muestral. No obstante, perturbaciones exógenas extremas como la crisis de la COVID-19 introducen una varianza desproporcionada que puede llegar a quebrar las correlaciones históricas entre las variables y el factor. Para verificar si la especificación seleccionada (**M04**) con tratamiento de interpolación ARIMA del periodo-1-covid logra una **estabilización paramétrica** robusta, se ejecuta un análisis recursivo mediante ventanas expansivas. Tomando como punto de anclaje inicial el escenario pre-covid (diciembre de 2019), el algoritmo reestima el modelo trimestre a trimestre hasta marzo de 2026. Los resultados de esta evolución y sus desviaciones acumuladas se presentan a continuación en la Figura 6.4.

Como se observa en el panel de la Figura 6.4, el modelo definitivo exhibe una robustez estructural frente al mayor shock de la historia económica reciente. El examen pormenorizado de las trayectorias de las cargas en la Tabla 6.4 revela los siguientes patrones de ajuste::

Tabla 6.4: Estabilidad de las cargas factoriales bajo estimación recursiva (M04).

.

Indicador	**Carga Base (Dic-2019)**	**Carga Shock (Jun-2020)**	**Carga Final (Mar-2026)**	**Desviación Máxima Acumulada**
Bloque Mensual				
(`PMI_MANU`)	0,3870	0,3650	0,3040	**0,0852** (Dic-23)
(`VGE_TOT`)	0,3250	0,3740	0,3780	0,0546 (Mar-24)
(`IPI`)	0,3160	0,3670	0,3580	0,0515 (Jun-20)
(`AFIL_SS`)	0,2410	0,2350	0,2780	0,0364 (Mar-26)
(`IMP_BIENES`)	0,2120	0,2470	0,2630	0,0619 (Mar-21)
(`EXP_BIENES`)	0,1970	0,2550	0,2560	0,0621 (Dic-23)
Bloque Trimestral				
(`PIB`)	0,0644	0,0577	0,0615	**0,0109** (Mar-20)
(`OCUPADOS`)	0,0529	0,0497	0,0428	0,0146 (Dic-20)

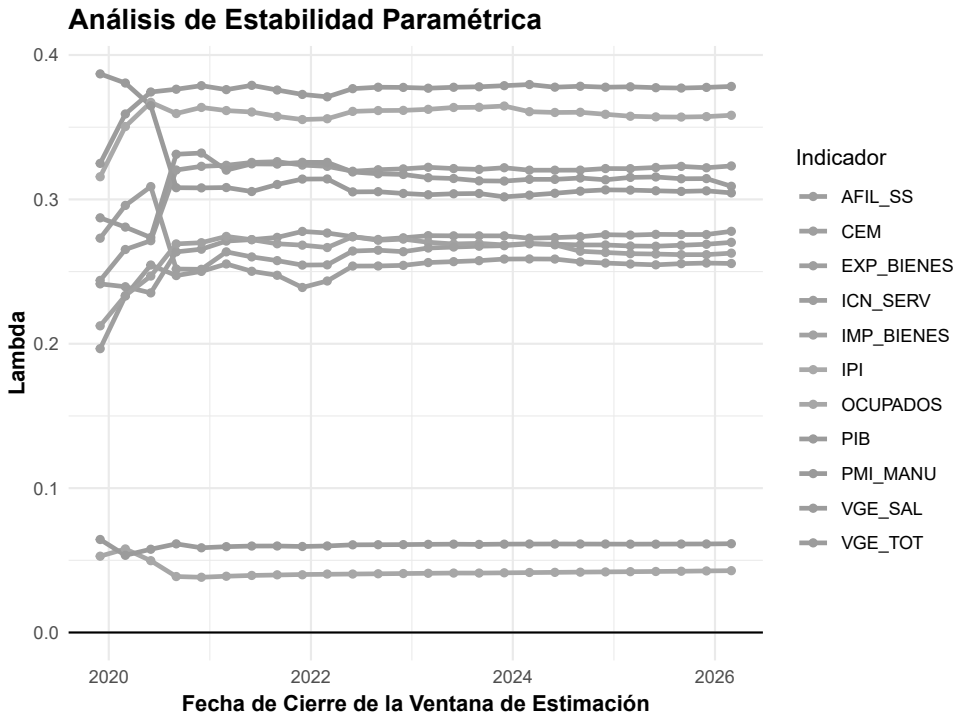
Nota: Elaboración propia.

En primer lugar, la desviación absoluta máxima de todo el modelo se registra en el PMI Manufacturero (`PMI_MANU`), alcanzando un pico de 0,0852 en diciembre de 2023 para estabilizarse en una carga final de 0,3040. Esta fluctuación demuestra que el algoritmo EM asimiló el ciclo post-confinamiento reajustando levemente el peso de la encuesta cualitativa, **pero sin afectar a su significatividad ni alterar la orientación del comovimiento estructural**.

En segundo lugar, las variables "duras" de actividad industrial y transaccional exhiben una consistencia simétrica excelente. El Índice de Producción Industrial (`IPI`) y las Ventas de Grandes Empresas (`VGE_TOT`) absorben la crisis con incrementos síncronos mínimos de 0,05 unidades en el peor trimestre del shock, manteniendo trayectorias planas y perfectamente paralelas durante el quinquenio posterior (2021–2026).

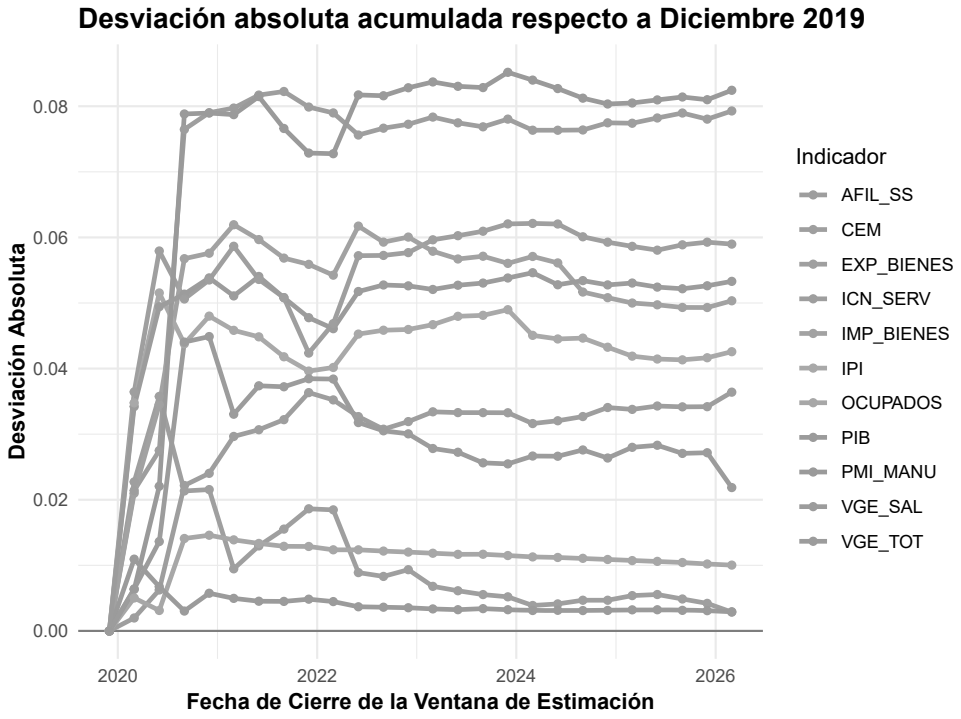
En tercer lugar, el anclaje de la variable objetivo es virtualmente perfecto. La carga del `PIB` se mantiene blindada en el entorno de los 0,06 puntos, registrando su oscilación máxima histórica en apenas un 0,0109 de desviación respecto a su base. Esto confirma que la relación interna que asocia el factor común mensual con el PIB real trimestral no sufrió ninguna distorsión estructural, garantizando la consistencia analítica de las actualizaciones predictivas del Filtro de Kalman.

Para contrastar la robustez del modelo elegido (**M04**), se presenta a continuación una estimación recursiva utilizando el panel de datos sin corregir la crisis del año 2020. Como se detalla en la Tabla 6.5, prescindir de la corrección univariante COVID-19 desestabiliza la estimación de los parámetros de



(a) Evolución dinámica de las cargas con Tratamiento

.



(b) Desviación absoluta acumulada respecto a diciembre de 2019

.

Figura 6.4: Análisis de estabilidad paramétrica en el Modelo M04 (Datos Tratados)

.

los parámetros del sistema a través de dos desviaciones:

Tabla 6.5: Distorsión de cargas factoriales en el escenario de datos con covid.

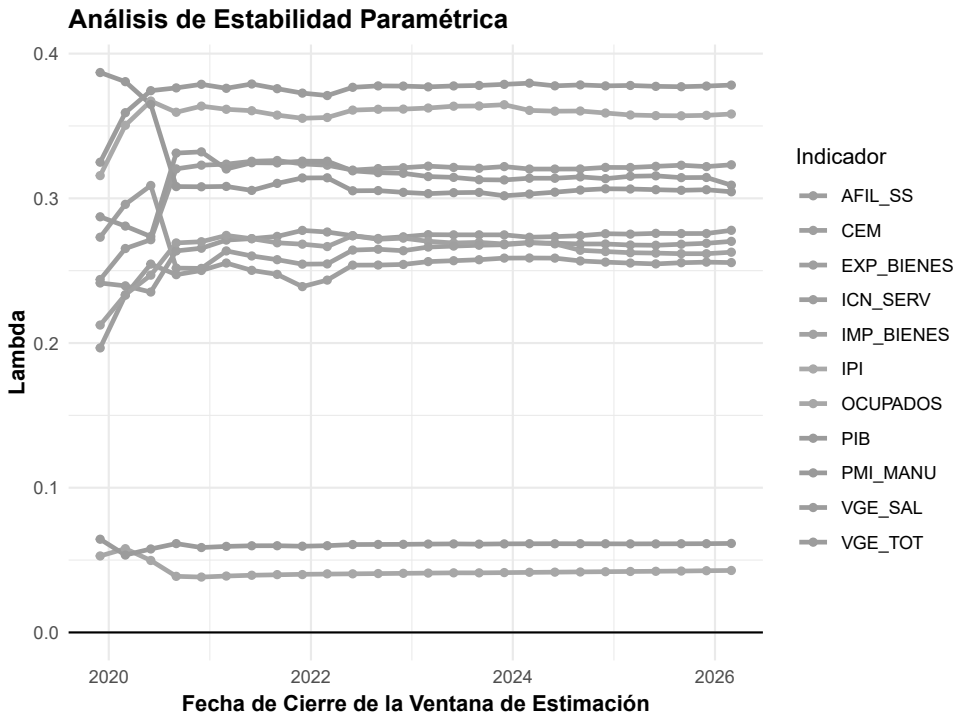
.

Indicador	**Carga Inicial**	**Pico del Shock (Jun-2020)**		**Escenario Final (Mar-2026)**	
	(Dic-2019)	**Sin Tratamiento**	**Con Tratamiento**	**Sin Tratamiento**	**Con Tratamiento**
`PMI_MANU`	0,3870	0,1590	0,3650	0,1620	0,3040
`IPI`	0,3160	0,4100	0,3670	0,4100	0,3580
`VGE_TOT`	0,3250	0,4090	0,3740	0,4120	0,3780
`AFIL_SS`	0,2410	0,3030	0,2350	0,3120	0,2780
`PIB`	0,0644	0,0933	0,0577	0,0846	0,0615

Ante la variabilidad anómala de las encuestas cualitativas, el modelo reduce de forma drástica y permanente la ponderación de `PMI_MANU`, cuya carga factorial disminuye en un 58 % (pasando de 0,3870 a 0,1620). Para compensar esta pérdida de información, el modelo incrementa el peso asignado a las variables cuantitativas como el `IPI` (+30 %) y la carga del propio `PIB` (+45 %). Esto provoca que el factor común latente altere su trayectoria para aproximarse a la caída puntual del −17,79 % observada en el PIB oficial.

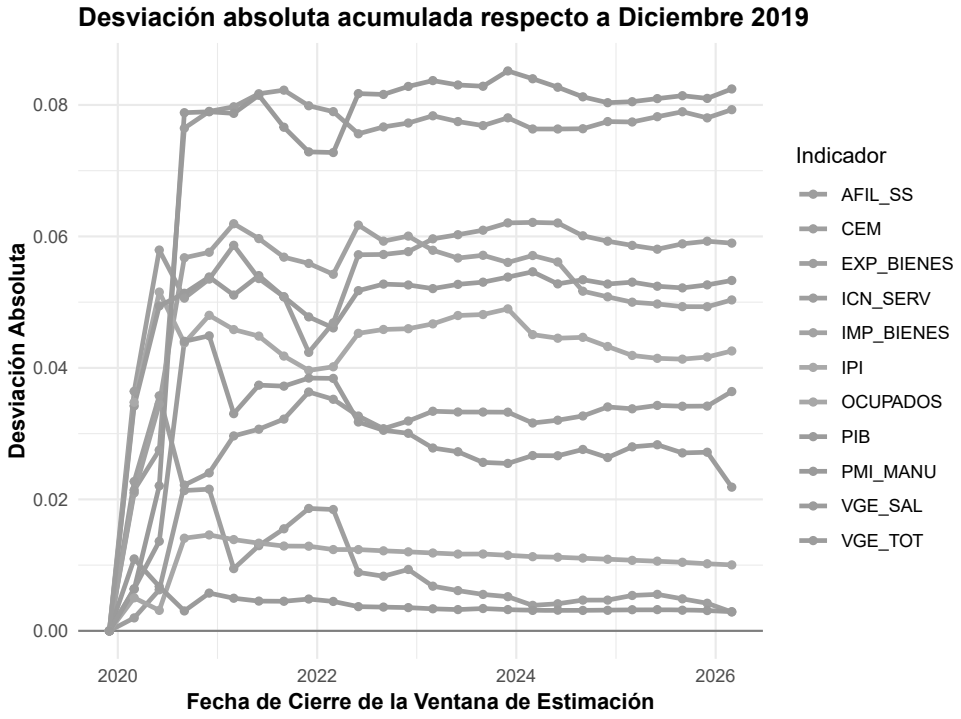
Por otro lado, la estimación sobre el panel sin tratamiento evidencia una marcada persistencia en la alteración de los parámetros. Una vez transcurrido el año 2020, las cargas factoriales no convergen hacia sus niveles de estabilidad histórica. En marzo de 2026, el coeficiente de PMI permanece en 0,1620 y el de IPI se mantiene en 0,4100. Este comportamiento demuestra que una observación atípica de varianza extrema perturba de forma permanente el proceso de optimización global de los parámetros del modelo. Las trayectorias resultantes de este escenario base y sus desviaciones acumuladas se recogen en el panel comparativo de la Figura 6.5.

Este contraste sirve de evidencia para observar como afecta el tratamiento con el resultado de garantizar la estabilidad histórica de las matrices del filtro y blindar la viabilidad del *nowcasting*.



(a) Evolución dinámica de las cargas sin tratamiento

.



(b) Desviación absoluta acumulada sin tratamiento

.

Figura 6.5: Análisis de inestabilidad paramétrica sin tratamiento de la COVID-19 (Datos Crudos)

.

6.2. Curva de evolución del *nowcast* intra-trimestral

La literatura especializada destaca que el valor añadido fundamental de un Modelo Factorial Dinámico no reside únicamente en la precisión de su estimación puntual, sino en su capacidad para actualizar de forma continua la proyección del PIB a medida que se difunden nuevos datos a lo largo del trimestre. Para cuantificar esta propiedad, se ha replicado el ejercicio de Giannone et al. (2008): se define un calendario cronológico de 31 pasos en la Tabla 6.6 que reproduce el orden real de publicación de las variables del panel en un trimestre cualesquiera (siempre que se respete el calendario de publicación de las instituciones) y se reestima el modelo de forma recursiva en cada paso, ocultando sistemáticamente toda observación no disponible en ese instante calendario, además, este calendario con los pasos de la llegada de nuevos datos servirá de apoyo para la siguiente sección a cerca del **Análisis de Descomposición de Noticias**.

Para cada uno de los trimestres de la muestra de validación (T1-2024 a T1-2026), se calcula el error absoluto de la predicción en cada uno de los 31 pasos y se agrega mediante la Raíz del Error Cuadrático Medio (RMSE). Este enfoque emula la curva de convergencia empleada por el *Federal Reserve Bank of New York* en su informe semanal *Nowcasting Report*. El resultado de esta métrica agregada se presenta en la Figura 6.6.

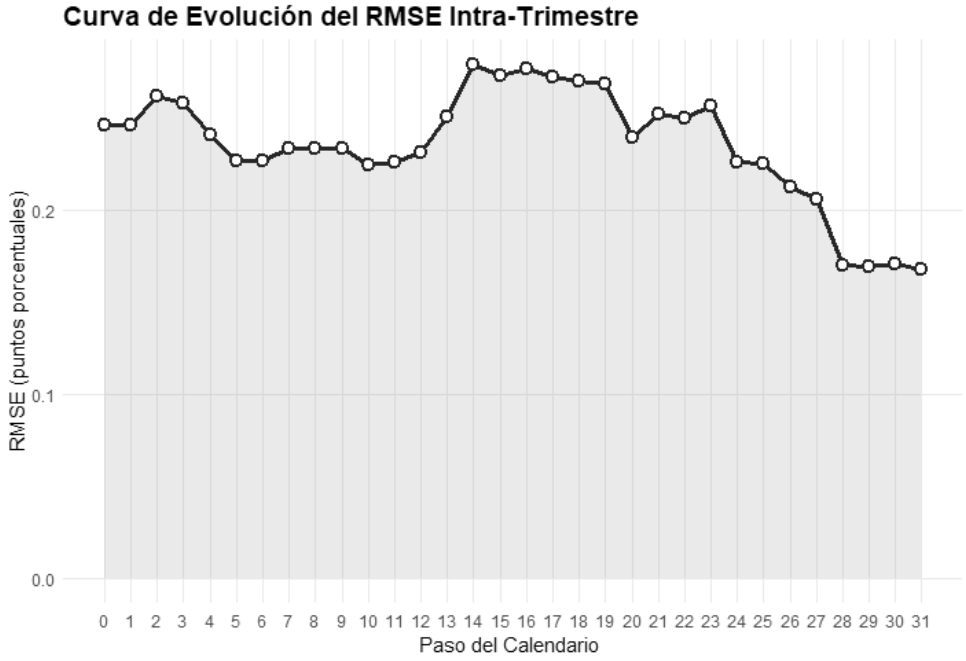


Figura 6.6: Curva de evolución del RMSE intra-trimestral, agregada sobre los trimestres T1:2024-T1:2026

.

A diferencia del patrón de caída monótona y lineal que se asume en la teoría, la curva del modelo revela una dinámica de absorción de información en dos fases. En la primera mitad del trimestre (pasos 0 al 16), el RMSE no experimenta reducciones significativas, e incluso exhibe un leve repunte oscilando en la franja de los 0.22 a 0.27 puntos porcentuales. Este fenómeno refleja el "ruido" inherente a los primeros compases del trimestre, donde los indicadores tempranos (*soft data*) a menudo emiten señales volátiles respecto a la inercia del PIB.

No obstante, a partir del paso 20, se observa un marcado efecto *Nowcasting Funnel* (embudo de convergencia). La reducción drástica del error, reduciéndose desde el entorno de los 0.25 hasta los

Tabla 6.6: Calendario cronológico de inyección de datos para el nowcasting (31 pasos).
.

Paso	Día aprox.	Variable(s)	Dato de referencia	Ejemplo T4 2025
Llegadas en el Mes 1 del trimestre (Ej: Octubre 2025)				
1	Día 2	AFIL_SS	Mes −1	Septiembre 2025
2	Día 3	PMI_MANU	Mes −1	Septiembre 2025
3	Día 6	IPI	Mes −2	Agosto 2025
4	Día 10	VGE_TOT, VGE_SAL	Mes −2	Agosto 2025
5	Día 15	CEM	Mes −1	Septiembre 2025
6	Día 15	ICN_SERV	Mes −2	Agosto 2025
7	Día 20	IMP_BIENES, EXP_BIENES	Mes −2	Agosto 2025
8	Día 28	OCUPADOS (EPA)	Trimestre −1	T3 2025
9	Día 30	PIB (Avance / Flash)	Trimestre −1	T3 2025
Llegadas en el Mes 2 del trimestre (Ej: Noviembre 2025)				
10	Día 2	AFIL_SS	Mes 1	Octubre 2025
11	Día 3	PMI_MANU	Mes 1	Octubre 2025
12	Día 6	IPI	Mes −1	Septiembre 2025
13	Día 10	VGE_TOT, VGE_SAL	Mes −1	Septiembre 2025
14	Día 15	CEM	Mes 1	Octubre 2025
15	Día 15	ICN_SERV	Mes −1	Septiembre 2025
16	Día 20	IMP_BIENES, EXP_BIENES	Mes −1	Septiembre 2025
Llegadas en el Mes 3 del trimestre (Ej: Diciembre 2025)				
17	Día 2	AFIL_SS	Mes 2	Noviembre 2025
18	Día 3	PMI_MANU	Mes 2	Noviembre 2025
19	Día 6	IPI	Mes 1	Octubre 2025
20	Día 10	VGE_TOT, VGE_SAL	Mes 1	Octubre 2025
21	Día 15	CEM	Mes 2	Noviembre 2025
22	Día 15	ICN_SERV	Mes 1	Octubre 2025
23	Día 20	IMP_BIENES, EXP_BIENES	Mes 1	Octubre 2025
Llegadas Post-Trimestre (Ej: Enero 2026)				
24	Día 2	AFIL_SS	Mes 3	Diciembre 2025
25	Día 3	PMI_MANU	Mes 3	Diciembre 2025
26	Día 6	IPI	Mes 2	Noviembre 2025
27	Día 10	VGE_TOT, VGE_SAL	Mes 2	Noviembre 2025
28	Día 15	CEM	Mes 3	Diciembre 2025
29	Día 15	ICN_SERV	Mes 2	Noviembre 2025
30	Día 20	IMP_BIENES, EXP_BIENES	Mes 2	Noviembre 2025
31	Día 28	OCUPADOS (EPA)	Trim. Actual	T4 2025

0.17 puntos porcentuales coincide geométricamente con la llegada de los indicadores duros (*hard data*) correspondientes al segundo y tercer mes del trimestre (tales como la Afiliación a la Seguridad Social, Ventas de Grandes Empresas o Consumo de Cemento). Estos resultados sugieren que la consolidación de la señal macroeconómica en España requiere atravesar el ecuador del trimestre, momento en el cual el Filtro de Kalman logra anclar la estimación latente purgando la volatilidad inicial.

6.3. Descomposición de noticias y la anatomía de las revisiones

Una de las funciones principales del *nowcasting* es comprender qué vectores de información están motivando cada revisión de la estimación inicial del PIB. Tal y como se formalizó teóricamente en la Sección 3.3.3.3, el marco de Bańbura y Modugno (2014) permite descomponer algebraicamente la revisión del *nowcast* como la suma de las innovaciones de los nuevos datos publicados, ponderadas por su respectiva Ganancia de Kalman. Para su implementación, se ha diseñado una función denominada *Descomposicion-noticias* que simula el flujo de información en pseudo tiempo real. Esta rutina llama internamente a la función `news()` del paquete `dfms`, la cual ejecuta el cálculo operando sobre el Vector de Estado Ampliado (`ss_full`). Matemáticamente, extrae las covarianzas cruzadas del suavizador ($P_1 P_2^{-1}$) para determinar el peso empírico de cada innovación (ν, o *news*) frente al ruido. Para ilustrar el proceso de actualización del modelo, se simula el flujo de información en pseudo-tiempo real durante el cuarto trimestre de 2025 (**T4 2025**). Partiendo de una estimación incondicional inicial, el goteo de publicaciones macroeconómicas expande iterativamente el conjunto de información disponible (Ω_v), permitiendo al Filtro de Kalman afinar la estimación del estado latente. La Figura 6.7 despliega esta trayectoria dinámica, desagregando además la contribución marginal por bloques estructurales.

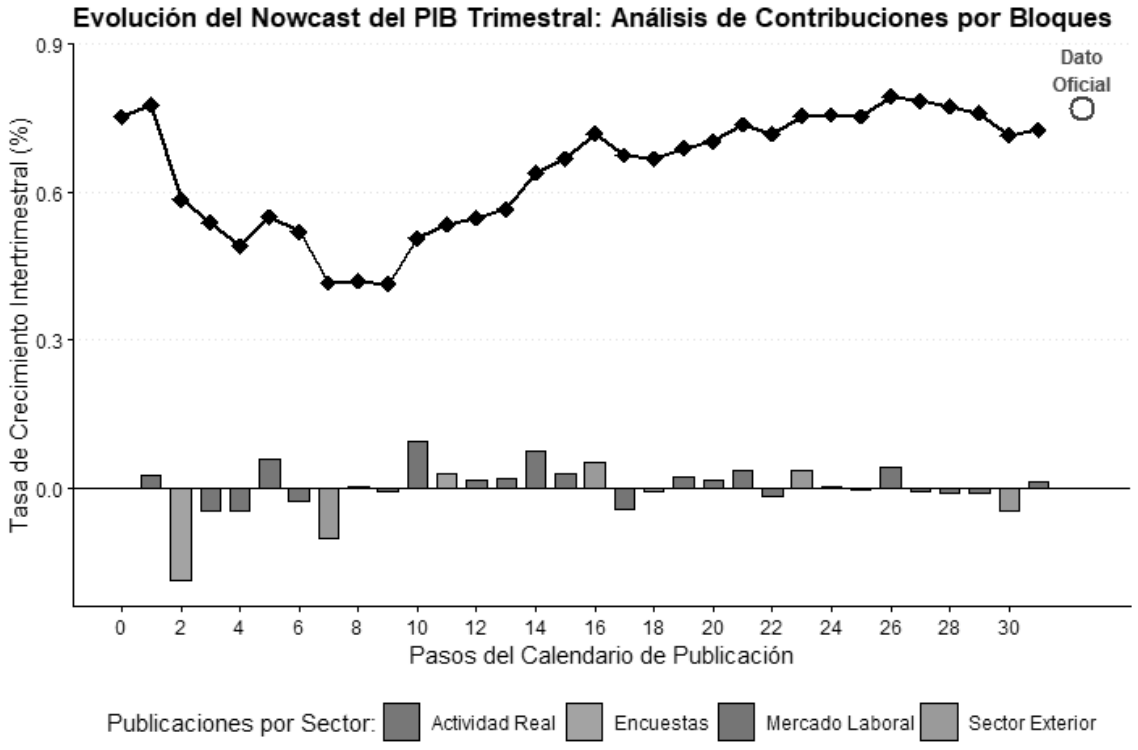


Figura 6.7: Evolución del nowcast intra-trimestral y contribución por bloques (T4-2025)

.

Como se evidencia en la Figura 6.7, la predicción no traza una línea monótona hacia el dato oficial, sino que existe la volatilidad característica del ejercicio de predicción en tiempo real, con subida y bajadas constantes. En los compases iniciales del trimestre, el modelo depende de los datos de encuestas (*soft data*), los cuales aportan direccionalidad temprana pero introducen cierto ruido. Sin embargo, a medida que avanza el calendario y se publican datos cuantitativos de actividad real y mercado laboral (*hard data*), el filtro reasigna pesos dinámicamente, anclando la predicción y propiciando una convergencia precisa hacia el dato oficial de Contabilidad Nacional, siendo el dato obtenido por el modelo en el paso 31 de 0.726 % frente al 0.77 % PIB oficial de primera estimación.

Mientras que la Figura 6.7 ofrece una visión más general de los datos que dirigen el modelo, la Figura 6.8 detalla el impacto marginal exacto de cada indicador en el momento exacto de su publicación.

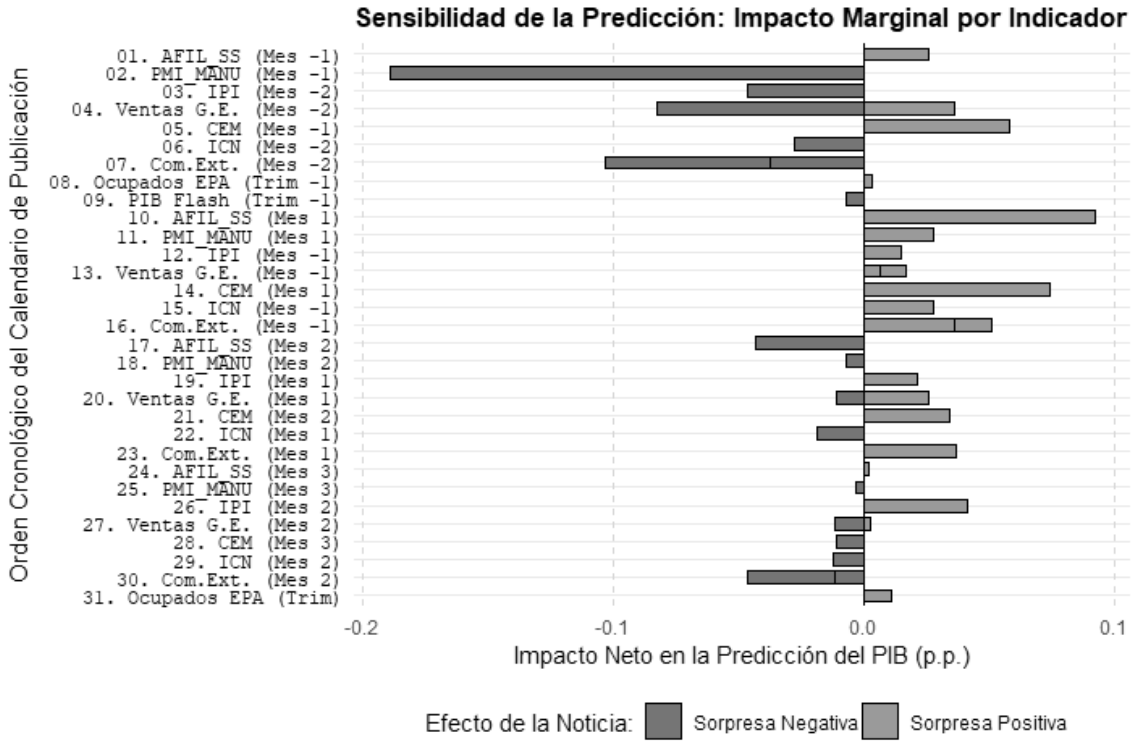


Figura 6.8: Sensibilidad del nowcast: Impacto marginal cronológico por indicador (T4-2025)

.

Esta descomposición ilustra el proceso de revisión secuencial de la estimación ante nuevos datos. Resulta crucial notar que una revisión de gran magnitud no implica necesariamente que el dato publicado sea excepcional, sino que el producto entre la innovación estadística y la Ganancia de Kalman ha sido elevado. Para comprender matemáticamente esta sensibilidad, la Tabla 6.7 desglosa la parametrización interna del filtro, revelando los pesos y las innovaciones subyacentes que determinan dichas correcciones.

La Tabla 6.7 presenta un desglose secuencial de la matriz de innovaciones del cuarto trimestre. Este ordenamiento cronológico permite observar como el Filtro de Kalman separa la señal subyacente del ruido estadístico a medida que se acumula nueva información:

- **El ajuste agresivo temprano:** En los primeros compases del proceso de actualización, las mayores desviaciones provienen típicamente de encuestas cualitativas. El Paso 2 muestra el primer sobresalto significativo: el PMI Manufacturero (Mes -1) reporta 51,5, lo que supone un fallo de

Tabla 6.7: Desglose trimestral paso a paso de la Ganancia e Impacto cronológico (T4 2025)

.

Paso	Mes/Trim	Indicador	Obs.	Pred.Kalman	ν	g	Imp. (p.p.)
1	AFIL_SS (Mes -1)	AFIL_SS	0.003	0.002	0.002	0.170	0.026
2	PMI_MANU (Mes -1)	PMI_MANU	51.500	54.005	-2.505	0.001	-0.189
3	IPI (Mes -2)	IPI	-0.002	0.007	-0.008	0.056	-0.047
4	Ventas G.E. (Mes -2)	VGE_TOT	-0.015	-0.001	-0.014	0.061	-0.083
4	Ventas G.E. (Mes -2)	VGE_SAL	0.009	0.005	0.005	0.078	0.036
5	CEM (Mes -1)	CEM	0.044	0.005	0.039	0.015	0.058
6	ICN (Mes -2)	ICN_SERV	-0.004	0.007	-0.011	0.026	-0.028
7	Com.Ext. (Mes -2)	EXP_BIENES	-0.059	0.007	-0.066	0.010	-0.066
7	Com.Ext. (Mes -2)	IMP_BIENES	-0.016	0.013	-0.029	0.013	-0.038
8	Ocupados EPA (Trim -1)	OCUPADOS	0.005	0.003	0.001	0.026	0.003
9	PIB Flash (Trim -1)	PIB	0.006	0.006	0.000	0.337	-0.007
10	AFIL_SS (Mes 1)	AFIL_SS	0.007	0.001	0.006	0.153	0.092
11	PMI_MANU (Mes 1)	PMI_MANU	52.100	51.649	0.451	0.001	0.028
12	IPI (Mes -1)	IPI	0.003	0.001	0.003	0.053	0.015
13	Ventas G.E. (Mes -1)	VGE_SAL	0.004	0.002	0.001	0.076	0.011
13	Ventas G.E. (Mes -1)	VGE_TOT	0.009	0.008	0.001	0.059	0.006
14	CEM (Mes 1)	CEM	0.045	-0.010	0.055	0.014	0.074
15	ICN (Mes -1)	ICN_SERV	0.016	0.006	0.011	0.026	0.027
16	Com.Ext. (Mes -1)	IMP_BIENES	0.036	0.008	0.028	0.013	0.036
16	Com.Ext. (Mes -1)	EXP_BIENES	0.045	0.030	0.015	0.010	0.015
17	AFIL_SS (Mes 2)	AFIL_SS	-0.002	0.002	-0.004	0.120	-0.043
18	PMI_MANU (Mes 2)	PMI_MANU	51.500	51.793	-0.293	0.000	-0.007
19	IPI (Mes 1)	IPI	0.006	0.002	0.004	0.059	0.021
20	Ventas G.E. (Mes 1)	VGE_TOT	0.005	0.001	0.004	0.065	0.025
20	Ventas G.E. (Mes 1)	VGE_SAL	0.003	0.005	-0.001	0.080	-0.011
21	CEM (Mes 2)	CEM	0.024	-0.007	0.030	0.011	0.034
22	ICN (Mes 1)	ICN_SERV	-0.004	0.003	-0.007	0.028	-0.019
23	Com.Ext. (Mes 1)	EXP_BIENES	0.029	-0.006	0.034	0.011	0.037
23	Com.Ext. (Mes 1)	IMP_BIENES	0.003	0.003	0.000	0.014	0.000
24	AFIL_SS (Mes 3)	AFIL_SS	0.002	0.002	0.000	0.075	0.002
25	PMI_MANU (Mes 3)	PMI_MANU	49.600	51.334	-1.734	0.000	-0.003
26	IPI (Mes 2)	IPI	0.011	0.003	0.008	0.049	0.041
27	Ventas G.E. (Mes 2)	VGE_TOT	0.005	0.004	0.001	0.054	0.002
27	Ventas G.E. (Mes 2)	VGE_SAL	0.003	0.005	-0.002	0.063	-0.012
28	CEM (Mes 3)	CEM	-0.019	-0.004	-0.015	0.007	-0.011
29	ICN (Mes 2)	ICN_SERV	0.004	0.010	-0.006	0.023	-0.013
30	Com.Ext. (Mes 2)	EXP_BIENES	-0.037	0.002	-0.039	0.009	-0.035
30	Com.Ext. (Mes 2)	IMP_BIENES	-0.003	0.008	-0.011	0.011	-0.012
31	Ocupados EPA (Trim)	OCUPADOS	0.008	0.004	0.004	0.026	0.011

predicción de $\nu = -2{,}505$. Debido a la varianza inherente a los indicadores de sentimiento, el algoritmo penaliza fuertemente este dato asignándole una ganancia marginal ($g = 0{,}0008$). Sin embargo, la magnitud bruta del error provoca una revisión severa de $-0{,}189$ puntos porcentuales en el *nowcast*. En el Paso 9, el modelo absorbe el error de predicción del propio PIB del trimestre anterior, generando una corrección ligera del $-0{,}007$ p.p., ya que la ganancia estructural de la variable objetivo es extremadamente alta ($g = 0{,}336$).

La estabilización mediante Hard Data (Meses 1 y 2): A partir del Paso 10, los indicadores del mercado laboral ajustan significativamente la estimación. El dato de Afiliación presenta una ligera innovación positiva ($\nu = +0{,}006$) pero, dado que este bloque ostenta una alta correlación contemporánea con el factor inobservable, la matriz de ganancias le otorga un peso elevado ($g = 0{,}1531$). Así, una variación minúscula se traduce en una revisión de $+0{,}092$ p.p. Esta dinámica de estabilización se repite a lo largo del trimestre con variables como el Consumo de Cemento o las Exportaciones.

- **Cierre de trimestre y reducción de la incertidumbre (Mes 3):** Conforme el modelo ingiere más información (Pasos 24 en adelante), el algoritmo converge y las actualizaciones en el *nowcast* pierden volatilidad. Por ejemplo, en el Paso 25 el PMI Manufacturero (Mes 3) vuelve a fallar marcadamente respecto a lo esperado ($\nu = -1{,}733$), pero la ganancia de Kalman asignada ha caído a $g = 0{,}000$; la información de este indicador a final de trimestre se asume como ruido frente a un panorama ya consolidado por los datos reales (*hard data*). Finalmente, el Paso 31 incorpora los Ocupados de la EPA, validando la inercia positiva trimestral y fijando la proyección final en $+0{,}011$ p.p. sobre el último ajuste.

Estos resultados indican que las revisiones del *nowcast* no dependen únicamente del tamaño absoluto del error de predicción en los indicadores mensuales. En la práctica, el Filtro de Kalman pondera cada innovación en función de su fiabilidad histórica y su relación señal-ruido, reasignando dinámicamente el peso de cada variable en la reconstrucción continua del ciclo económico.

6.4. Análisis de sensibilidad

Para completar la evaluación predictiva, se somete el modelo a un análisis de sensibilidad. La literatura clásica de series temporales estructurales (Harvey, 1989) advierte sobre la necesidad de tratar explícitamente los cambios de régimen para no sesgar la estimación del estado latente. En esta sección se insertan distintos shocks puntuales al modelo seleccionado con el fin de comprobar si su respuesta es proporcional y económicamente interpretable.

6.4.1. Diseño de la prueba

Para cada uno de cuatro escenarios se inserta un *shock* en la última observación disponible del 4T de 2025, expresado en unidades de su desviación típica histórica, y se reestima el MFD completo (incluyendo $\hat{\Lambda}$, $\hat{A}$, $\hat{Q}$ y $\hat{R}$) con dicho valor alterado dentro de la muestra. A diferencia de un ejercicio de sensibilidad con parámetros fijos –que aislaría únicamente el efecto sobre el estado filtrado–, esta reestimación completa permite responder a una pregunta más realista: si la perturbación, de haber ocurrido, también desestabilizaría la estructura factorial estimada.

Cada escenario se evalúa en dos magnitudes –una moderada (en torno a $1\hat{\sigma}$) y otra ampliada (entre $2\hat{\sigma}$ y $4\hat{\sigma}$, con un factor de escalado distinto según la variable)– con el fin de contrastar si la respuesta del modelo escala de forma aproximadamente proporcional a la magnitud del *shock*, o si por el contrario exhibe algún umbral de inestabilidad o comportamiento no lineal dentro de este rango. Los cuatro escenarios considerados son: **Expansión** y **Recesión** (shocks simétricos y simultáneos sobre PMI de manufacturas, afiliación a la Seguridad Social, IPI, ventas de grandes empresas y consumo de cemento), **Shock exterior** (contracción de exportaciones e importaciones de bienes junto con el PMI de manufacturas) y **Mercado laboral** (caída conjunta de afiliación y ocupados EPA).

La Tabla 6.8 presenta, para cada escenario, la variación del *nowcast* del PIB respecto al escenario Base (0,818 %) bajo ambas magnitudes de *shock*, junto con el ratio de escalado observado –el cociente entre el efecto de la magnitud ampliada y la moderada– y el factor de escalado que cabría esperar bajo el supuesto de proporcionalidad estricta entre la magnitud del *shock* y su efecto sobre el *nowcast*.

Tabla 6.8: Sensibilidad del *nowcast* del PIB ante dos magnitudes de *shock*, por escenario.

.

Escenario	**Δ moderado** (p.p.)	**Δ ampliado** (p.p.)	**Ratio observado**	**Ratio esperado** (lineal)	**Desviación**
Expansión	+0,154	+0,335	2,18	2,00	+9 %
Desaceleración	−0,136	−0,264	1,94	2,00	−3 %
Shock exterior	−0,062	−0,180	2,90	3,22	−10 %
Mercado laboral	−0,020	−0,055	2,75	2,75	0 %

En los cuatro escenarios, el ratio de escalado observado se sitúa dentro de un margen de ±10 % respecto al que resultaría de una respuesta perfectamente proporcional, sin evidencia de cambios de régimen dentro del rango de magnitudes evaluado ($1\hat{\sigma}$ a $4\hat{\sigma}$). El caso de **Mercado laboral** resulta más ilustrativo: el ratio observado (2,75) coincide de forma exacta con el factor de escalado teórico del propio diseño del escenario. La Figura 6.9 y la Figura 6.10 ilustran, respectivamente, la magnitud de estas variaciones y su patrón de escalado.

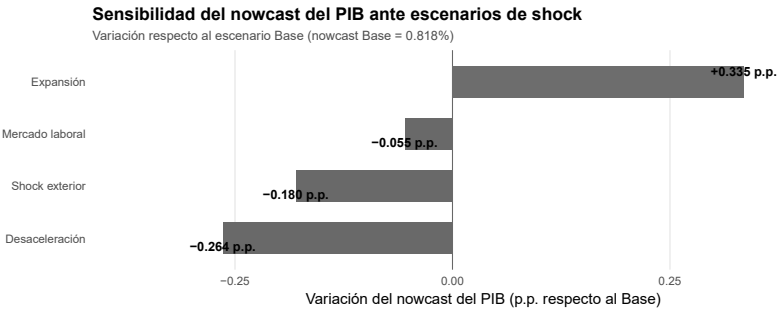


Figura 6.9: Variación del *nowcast* del PIB por escenario (magnitud ampliada), respecto al escenario Base

.

6.4.2. Estabilidad de las cargas factoriales

Además de observar como afecta sobre el *nowcast* puntual de un trimestre, interesa examinar si la introducción del *shock* dentro de la muestra de estimación desestabiliza la estructura factorial del modelo, esto es, si $\hat{\Lambda}$ se reestima de forma sustancialmente distinta al incorporar la observación atípica. La Figura 6.11 representa, para cada escenario y variable, la carga factorial del modelo Base frente a la obtenida tras la reestimación con el *shock* introducido en la stimación de las mismas.

Para la generalidad de las variables del panel, la variación de $\hat{\lambda}_i$ es reducida (inferior al 2 % en valor relativo) y escala de forma aproximadamente proporcional a la magnitud del *shock*. Se detecta, no obstante, una excepción relevante: la carga factorial de **OCUPADOS** en el escenario de Mercado laboral se desestabiliza de forma supra-proporcional, pasando de una variación del −0,27 % bajo el *shock* moderado a un −2,17 % bajo el ampliado –un ratio de 8,04, muy superior al factor de escalado

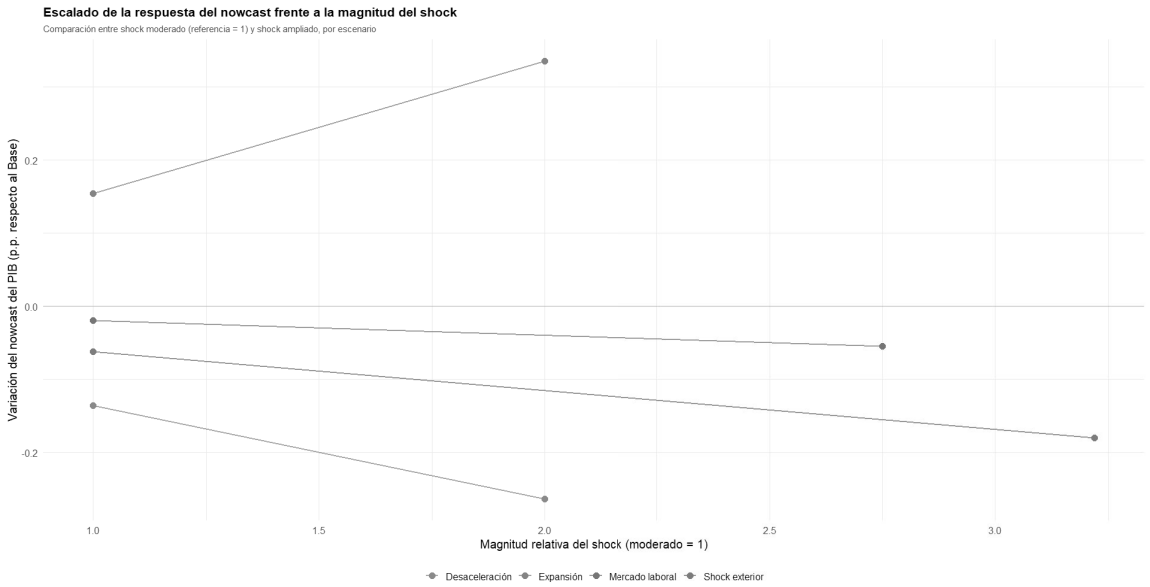


Figura 6.10: Escalado de la respuesta del *nowcast* frente a la magnitud del *shock*, por escenario

.

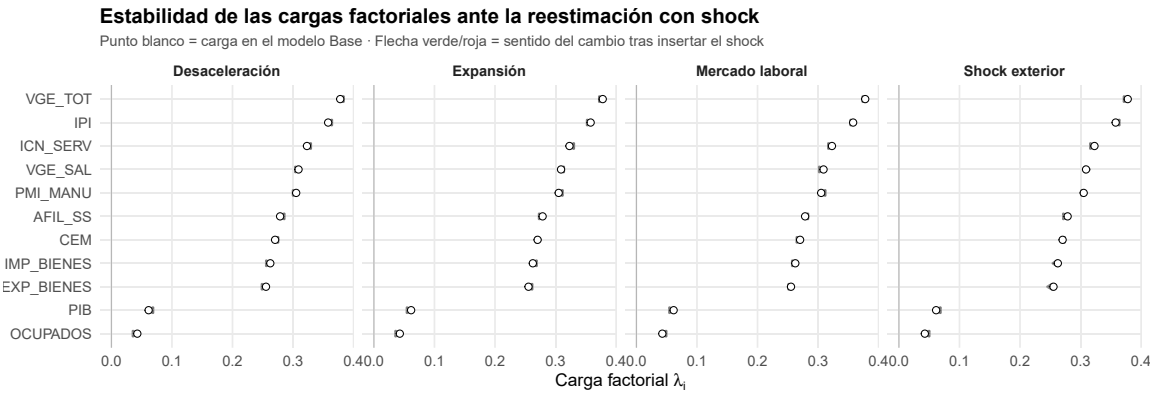


Figura 6.11: Movimiento de las cargas factoriales (λ_i) tras la reestimación con *shock*, por escenario: desaceleración, expansión, mercado laboral y shock exterior

.

esperado de 2,75 para ese escenario–. Esta anomalía es coherente con una limitación estructural ya que al ser *OCUPADOS* una serie de frecuencia trimestral cuya historia disponible arranca en 2005 ($N = 84$ observaciones efectivas, frente a las 303-375 del resto del panel), su carga factorial se estima con un número sensiblemente menor de grados de libertad, lo que la hace más sensible a la inserción de valores atípicos de magnitud elevada –precisamente la variable directamente perturbada en este escenario–.

Cabe señalar que algunas de las variaciones porcentuales más extremas observadas en la comparación de robustez (por ejemplo, en *AFIL_SS* bajo el escenario de Mercado laboral) no reflejan una inestabilidad real, sino un artefacto numérico. Al partir de una variación porcentual muy próxima a cero, cualquier cociente calculado sobre ella tiende a distorsionarse, sin que ello implique un movimiento considerable de la carga factorial en términos absolutos.

El análisis de sensibilidad demuestra la robustez numérica de la especificación seleccionada (M04) frente a perturbaciones puntuales de magnitud moderada y elevada: la respuesta del *nowcast* escala de forma aproximadamente lineal con la magnitud del *shock*, sin manifestar puntos de quiebre dentro del rango evaluado, y la estructura factorial se mantiene mayoritariamente estable, con la única excepción relevante concentrada en la variable con menor profundidad muestral del panel (*OCUPADOS*). Este resultado es coherente con el principio de parsimonia: un modelo correctamente especificado no debería amplificar de forma desproporcionada perturbaciones de cierta magnitud, en contraste con el comportamiento desarrollado para el período de la COVID-19 en la Sección 5.2, donde la perturbación de naturaleza común y no propia sí exigió un tratamiento diferenciado.

Capítulo 7

Implementación tecnológica y creación de la aplicación

El desarrollo y la estimación de los modelos estudiados en los capítulos anteriores se ha llevado a cabo siempre en R. Sin embargo, uno de los problemas que surgen en la práctica es que el usuario final que usa los datos del modelo para sus informes o presentaciones no tienen por qué ser conocedores de este lenguaje de programación. Como consecuencia, se genera una dependencia hacia la persona que ha programado el modelo, restando agilidad e impacto a la herramienta en el día a día de la entidad. Como solución a este inconveniente, se ha decidido construir una aplicación interactiva. De esta manera, cualquier usuario, sin necesidad de saber programar en R, puede acceder a la herramienta de *nowcasting*, consultar las predicciones para el trimestre actual, visualizar el estado del factor común e interactuar con los indicadores y el modelo. Esto ha sido posible gracias a la librería Shiny, cuyo funcionamiento y arquitectura se describen a lo largo de esta sección.

Shiny es una librería de R creada específicamente para desarrollar aplicaciones web interactivas que permiten al usuario manipular datos y ejecutar funciones de R sin necesidad de interactuar directamente con el código con una interfaz de usuario amigable. Una de sus grandes ventajas es que para construir estas aplicaciones no son necesarios conocimientos avanzados de HTML, CSS o JavaScript, que son los lenguajes habituales de programación web.

El objetivo de esta sección no es crear un manual sobre programación, pero sí proporcionar una comprensión clara de la arquitectura y el esquema de construcción de este tipo de herramientas para finalizar con la presentación de las funcionalidades más atractivas de la aplicación creada para explotar el modelo del presente proyecto. En la Sección 7.1 se describe la estructura fundamental de cualquier aplicación Shiny, mientras que en la Sección 7.3 se detalla el diálogo interno entre sus objetos para lograr que la aplicación funcione.

7.1. Introducción y arquitectura

Shiny basa su funcionamiento en un esquema de programación reactiva. En este tipo de programación, los valores pueden cambiar a lo largo del tiempo y existen expresiones que "reaccionan" automáticamente a esos cambios. Así, se vinculan de forma dinámica los objetos de entrada (*inputs*) con los de salida (*outputs*), provocando que, cada vez que el usuario hace una modificación, los resultados se actualicen en tiempo real. Toda aplicación Shiny necesita dos elementos fundamentales para ser operativa: una página web que muestra la interfaz al usuario, y un ordenador (o servidor) que esté ejecutando R por detrás. El esquema de funcionamiento se divide estrictamente en dos partes que interactúan entre sí: la *UI* y el *Server*, esta estructura se puede visualizar en la Figura 7.1. Diferenciarlas es el paso fundamental para entender cualquier aplicación de este tipo:

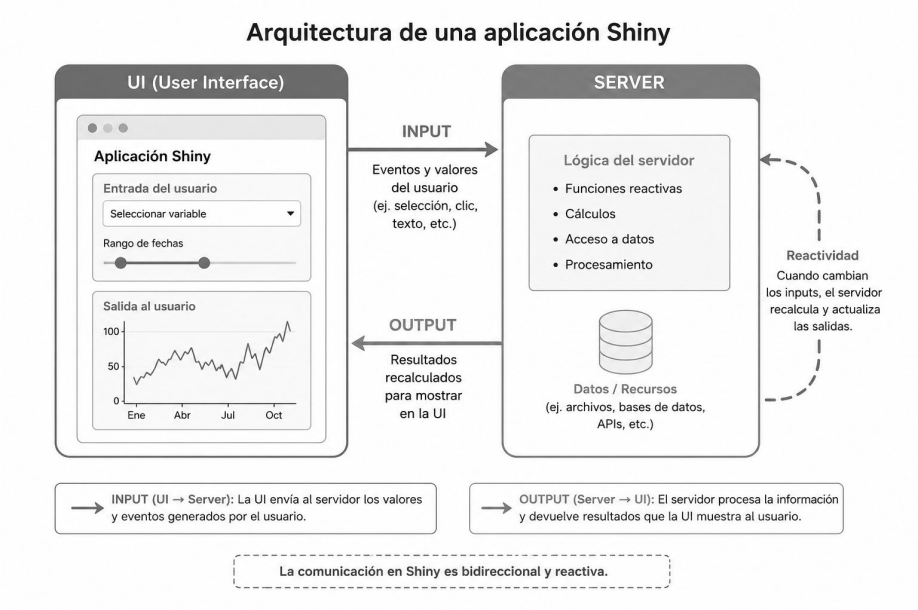


Figura 7.1: Diagrama básico de la arquitectura y reactividad de una aplicación Shiny.

.

- **UI (Interfaz gráfica):** En esta parte se describe el aspecto, el diseño y la disposición de la aplicación. Es la encargada de alojar los controles para recibir las peticiones del usuario (*inputs*) y de reservar los espacios donde se mostrarán los gráficos o tablas (*outputs*).

- **Server (Código del servidor):** Es el "cerebro" de la aplicación. Aquí se ejecuta el trabajo estadístico necesario (en nuestro caso, el Filtro de Kalman o la estimación factorial). Se utiliza código de R tradicional, con la diferencia de que los parámetros numéricos o selecciones del usuario entran como variables dinámicas. Como resultado, se generan elementos reactivos que se devuelven a la interfaz.

A la hora de programar la herramienta, se puede optar por alojar ambas partes en un único archivo, o separarlas en dos archivos distintos (`ui.R` y `server.R`) para mantener el código más organizado en proyectos complejos. En el caso concreto de este proyecto se ha decidido alojar ambas partes en un único archivo pues encuentro más organizado esta manera y me permite no olvidarme de cruzar siempre el elemento que añado en la UI con el Server en mi código. Para ilustrar el funcionamiento real de la aplicación, es necesario comprender cómo se articulan los diferentes objetos desde que el usuario hace una petición hasta que R devuelve el resultado. Este ciclo consta de tres fases principales: la lectura de *inputs*, la construcción de elementos reactivos y la visualización de los *outputs*.

7.1.1. Lectura de Inputs

Shiny dispone de una gran variedad de *widgets* o controles web para que el usuario introduzca información. Todos ellos requieren, como mínimo, dos argumentos básicos: un identificador interno (`inputId`) para poder llamar a ese valor desde el servidor, y una etiqueta textual (`label`) que es la que leerá el usuario en la pantalla. A través de la Figura 7.2 se pueden ver algunos casos de uso y ejemplos.

Para utilizar el valor recogido por uno de estos *widgets* dentro de la parte del servidor, simplemente se emplea el comando `input$inputId`.

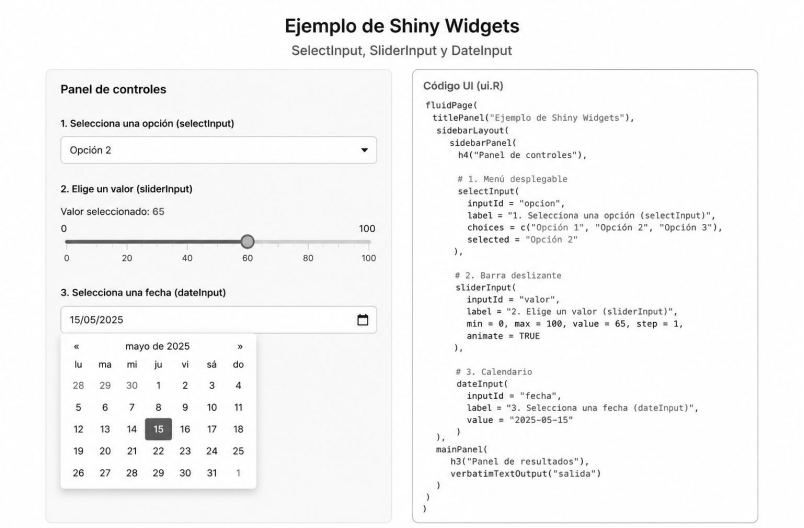


Figura 7.2: Ejemplos de varios *widgets* de entrada de Shiny

.

Siguiendo con los elementos reactivos, son objetos generados en la parte del servidor mediante funciones de R, utilizando la información que proviene de los *inputs*. Si la intención es que este elemento se envíe a la interfaz para que el usuario lo vea (por ejemplo, la figura de las cargas factoriales de mi modelo, se debe construir utilizando una función de la familia `render`.

La sintaxis genérica es la siguiente:

```
output$nombre_del_grafico <- renderPlot({
  # Codigo R necesario para crear el grafico
  # utilizando variables como input$inputId
})
```

Existen diversas funciones de este tipo dependiendo de lo que se quiera mostrar. Por ejemplo, `renderPlot` genera gráficos, `renderTable` construye tablas, y `renderText` devuelve variables numéricas o cadenas de texto.

7.1.2. Visualización de los Outputs

Una vez que el servidor ha construido el elemento reactivo y lo ha guardado en el objeto `output`, es necesario indicarle a la interfaz (UI) dónde y cómo debe dibujarlo. Para ello, se utilizan funciones de la familia `Output`, las cuales están emparejadas con las funciones `render` que se utilizaron en el paso anterior.

De este modo, si en el servidor creamos un gráfico con `renderPlot`, en la interfaz debemos reservar el hueco utilizando la función `plotOutput("nombre_del_grafico")`.

7.2. Personalización y librerías complementarias

Cabe resaltar que a pesar de que Shiny simplifica el uso de HTML para la ingeniería web, en la construcción de la aplicación de este trabajo se ha optado por enriquecer la herramienta utilizando librerías adicionales del ecosistema de R. Esto permite lograr un acabado más profesional y atractivo para el usuario final.

Entre las librerías implementadas destacan: Entre las librerías implementadas destacan:

- **shinydashboard** (Chang y Ribeiro, 2021): Permite abandonar el formato de página web tradicional para construir un cuadro de mando (*dashboard*). Proporciona un menú de navegación lateral, barras superiores y cajas de resumen, muy valoradas para la presentación de resultados.
- **plotly / dygraphs** (facing, Allaire, Owen, Gromer, y Thieurmel, 2018; Sievert, 2020): Se utilizan para transformar los gráficos estáticos de R en gráficos interactivos. Gracias a ellas, el usuario puede pasar el ratón por encima del gráfico para consultar el valor exacto de un mes concreto, hacer *zoom* en determinados periodos de interés mediante deslizadores temporales o aislar bloques sectoriales con un solo clic en la leyenda.
- **DT (DataTables)** (Xie, Cheng, y Tan, 2023): Implementa el renderizado de las matrices de datos bajo un diseño editorial institucional. Permite la ordenación dinámica, búsquedas en vivo y la manipulación de celdas interactivas, lo que viabiliza funciones avanzadas como la sincronización cruzada entre el clic de un gráfico y la iluminación automática de la celda correspondiente.
- **shinyjs / shinyWidgets** (Attali, 2021; Perrier, Meyer, y Granjon, 2023): Aportan el control y la sofisticación UX necesarios en entornos corporativos. *shinyjs* permite inyectar funciones de JavaScript para congelar o activar controles de la interfaz durante peticiones pesadas al servidor, mientras que *shinyWidgets* enriquece el conjunto de botones y los selectores con búsquedas predictivas integradas, optimizando el espacio visual de la aplicación.
- **ellmer / shinychat** (Schloerke y Cheng, 2024; Wickham, 2024): Constituyen la infraestructura de Inteligencia Artificial generativa local de la plataforma. Permiten anclar modelos de lenguaje avanzados (LLMs) al servidor de la aplicación para desplegar el chat interactivo de consulta econométrica y el módulo de redacción automática de notas de prensa en formato Markdown.
- **promises** (Cheng, 2023): Introduce la arquitectura de computación asíncrona distribuida. Su implementación es vital, ya que desvía los procesos de inferencia pesados de la IA o las descargas automáticas por API hacia hilos secundarios del procesador, impidiendo el bloqueo o la congelación de la interfaz gráfica de usuario.

La combinación de estos paquetes permite crear una aplicación de un nivel empresarial muy alto con gran interactividad y con un apoyo gráfico muy potente para convertir un modelo matemático abstracto en una herramienta interactiva.

7.3. Implantación de la aplicación y entorno de producción

La transformación del modelo en una plataforma digital se estructuró bajo un entorno integrado de producción orientado a la explotación interactiva. El objetivo de la herramienta es actuar como un puente de comunicación directa entre las estimaciones del modelo y la toma de decisiones ejecutivas de la entidad, agilizando la generación de información periódica de alta frecuencia demandada en los comités de dirección. Para ello, se dota al sistema de una interfaz reactiva que procesa el panel macroeconómico en tiempo real, aislando al analista de la complejidad algorítmica pero garantizando un control riguroso sobre cada etapa del proceso de predicción.

En la primera de las secciones de la plataforma, cuyo entorno visual se expone en la Figura 7.3, se despliega el cuadro de mando macroeconómico dedicado al Análisis Exploratorio de las series que se han seleccionado. Para obtener los datos actualizados al instante actual se ha programado un botón que al pulsarse comienza la tarea de descarga automática de las series cuyo proceso se ha detallado en la Sección 4.2. Una vez cargado el panel de datos, el usuario puede navegar de forma interactiva para monitorizar las alertas tempranas y la calidad del dato transversal.

Para continuar, la siguiente pestaña de la aplicación web aborda la estimación del factor central y el mapeo de cargas factoriales. Una vez realizado todo el preprocesado de las series, el servidor ejecuta el algoritmo de Banbura y Modugno, restringiendo el cómputo al panel óptimo central de once variables

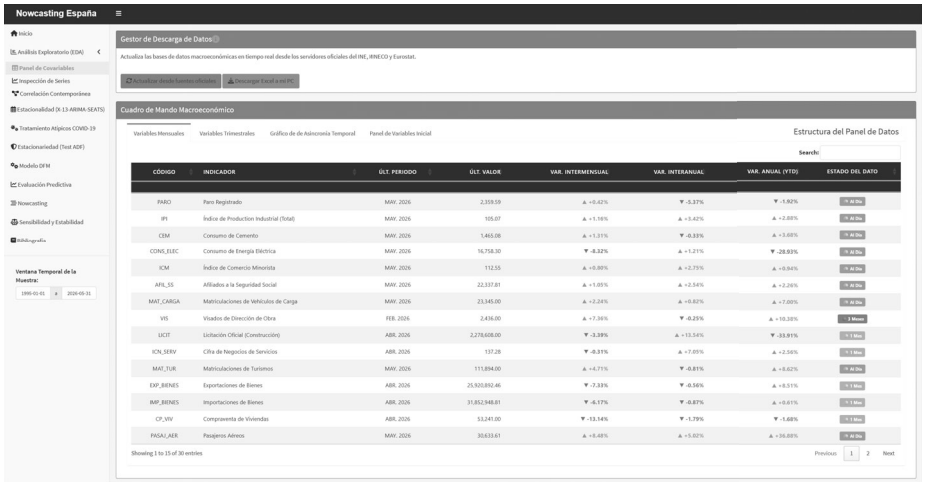


Figura 7.3: Interfaz del Cuadro de Mando Macroeconómico.

.

seleccionado en la investigación. Para poder representarse de forma conjunta en el mismo gráfico se agrega el factor mensual a frecuencia trimestral a través de la ponderación de Mariano y Murosawa, tal y como se aprecia en el mapeo de resultados de la Figura 7.4. De este modo, se evita el aplanamiento espurio del factor latente durante los trimestres de crecimiento ordinario, preservando la resolución cíclica.

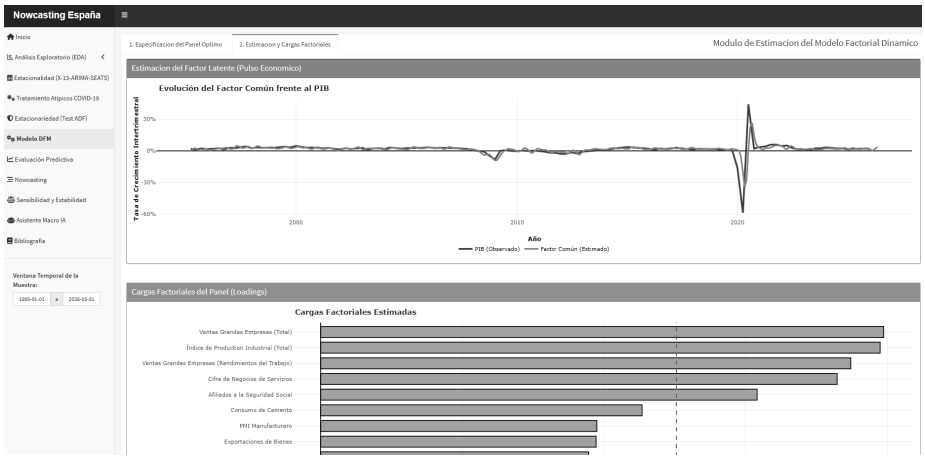


Figura 7.4: Visualización del Pulso Económico Estimado frente al PIB Observado.

.

Por su parte, el módulo relativo al algoritmo cronológico de Nowcasting por días hábiles representa el núcleo operativo de la plataforma en el día a día de la entidad, al automatizar el proceso de inyección secuencial de información fuera de la muestra. A diferencia de los simuladores estáticos, la aplicación implementa un planificador cronológico dinámico que resuelve de forma ordenada el calendario con retraso de los organismos públicos españoles (INE, MINECO y Agencia Tributaria). El servidor incorpora una función correctora de días hábiles que calcula la fecha exacta teórica de publicación de cada indicador para el año y trimestre seleccionado por el usuario, determinando el impacto marginal neto provocado por la sorpresa macroeconómica de cada publicación, tal y como se

Figura 7.5: Módulo de Nowcasting Automatizado y Descomposición Marginal de Noticias.

.

desglosa en el panel de la Figura 7.5.

Con el fin de dotar a la plataforma de la mayor automatización posible, se ha diseñado e implantado un ecosistema conversacional local basado en la integración de las librerías `ellmer` y `shinychat` que se observa su estructura en la Figura 7.6. Este módulo actúa como un consultor econométrico previamente entrenado que asiste al usuario en la interpretación de los resultados del modelo antes de un consejo de administración o comité. La arquitectura técnica se divide en dos motores de lenguaje que corren en local a través del servidor de Ollama, eliminando cualquier vulnerabilidad de filtración de datos o dependencia de conexión exterior, algo fundamental en un banco. **Asistente de Coyuntura (Llama 3.2):** Gestiona la interfaz del chat interactivo. El servidor enriquece el prompt del usuario de forma invisible inyectando las variables de estado reactivas del modelo actualizadas en ese preciso instante en la memoria RAM (como el Nowcast final calculado o el número de hitos integrados). Esto blinda al LLM contra alucinaciones de datos y ancla sus respuestas estrictamente a la realidad numérica de la economía española.

Figura 7.6: Ecosistema de Inteligencia Artificial Macroeconómica Local: Interfaz Conversacional de Consulta.

.

Capítulo 8

Conclusiones y líneas futuras de investigación

Tras examinar y desglosar los distintos parámetros, enfoques y técnicas analizadas, los resultados obtenidos confirman que es posible construir un marco formal y sistemático para estimar la actividad económica española en tiempo real a través de un Modelo Factorial Dinámico. La filosofía intrínseca al modelo de factores dinámicos permite resolver los problemas de convergencia y la explosión de parámetros de los métodos tradicionales, condensando la información de un gran panel de indicadores mensuales dentro de una representación de espacio de estados parsimoniosa. Para la entrada de datos en el Modelo Factorial Dinámico, se aplicó un flujo de trabajo estructurado en varias fases: la primera, corrección de la estacionalidad con el algoritmo X-13ARIMA-SEATS, la transformación logarítmica y en diferencias para asegurar la condición de estacionariedad débil y la estandarización final de la matriz.

La validez de este enfoque de reducción de la dimensión queda patente en los resultados obtenidos ante observaciones que el modelo no conocía, validándolo para la ventana pos-COVID (2024–2025), donde el modelo M04 registró una raíz del error cuadrático medio ($RMSE$) de 0,155 y un error absoluto medio (MAE) de 0,120 en una estructura de la economía similar a la actual, con la pandemia de la COVID-19 ya lejana. Estas métricas no solo mejoran el comportamiento de la inercia histórica univariante, un AR(1), sino que están en línea con los valores publicados por organismos de referencia como la AIReF, alcanzando un índice de concordancia perfecto (1,000) al anticipar la dirección de los puntos de giro del ciclo económico. Asimismo, la evaluación mediante la regresión de Mincer-Zarnowitz confirmó que las proyecciones carecen de sesgo sistemático y consiguen explicar de forma directa el 62,4 % de la varianza del crecimiento intertrimestral del Producto Interior Bruto fuera de la muestra.

En cuanto a los sectores y variables que aportan mayor contenido informativo para la revisión de las perspectivas de crecimiento, la criba exhaustiva determinó que el núcleo duro de la economía real (mercado laboral, producción física y demanda agregada) constituye la combinación estructural con mayor capacidad predictiva y estabilidad paramétrica de las cargas factoriales. Finalmente, cabe destacar que hubo que llevar a cabo un tratamiento *ad hoc* para mitigar el shock exógeno producido por la parada de la actividad a raíz del confinamiento inicial de la pandemia. Ante esta situación existió la necesidad de realizar un tratamiento posterior para que la entrada de volatilidad extrema no contaminara las relaciones estructurales históricas entre las variables que entran al MFD, por lo que un shock de una magnitud similar a la de la pandemia podría generar la necesidad de una nueva redefinición total del modelo como ocurrió con algunos MFD institucionales como el de la FED de NY o el nombrado anteriormente MiPred de la AIReF.

Apéndice A

Fundamentos Analíticos de Series Temporales

Este apéndice compendia de forma exhaustiva las definiciones formales, los teoremas estructurales y la formulación paramétrica que sustentan la metodología de series temporales univariantes y multivariantes aplicada en el presente trabajo. El desarrollo teórico se basa en las obras de referencia clásicas de Box et al. (2015), Hamilton (1994) y Peña (2010), en los fundamentos docentes de Aneiros (2025), así como en la vertiente computacional y de aplicación en el lenguaje de programación R tratada por Cryer y Chan (2008) y Cowpertwait y Metcalfe (2009).

A.1. Procesos Estocásticos y Operadores de Transformación

Definición A.1. *Un **proceso estocástico** se define como una familia de variables aleatorias* $\{X(t,\omega) : t \in \mathcal{T}, \omega \in \Omega\}$ *definidas sobre un espacio de probabilidad común* $(\Omega, \mathcal{F}, P)$.

Para cada instante $t \in \mathcal{T}$ *fijo,* $X(t,\cdot)$ *es una variable aleatoria definida sobre el espacio muestral* Ω. *Inversamente, para cada resultado elemental* $\omega \in \Omega$ *(fijo), la aplicación* $X(\cdot,\omega)$ *se denomina **realización**, **trayectoria** o **serie temporal** del proceso sobre el conjunto de índices* $\mathcal{T}$.

Definición A.2. *Se define una **serie temporal** como una realización parcial y finita de un proceso estocástico* $\{X_t\}_{t\in\mathbb{Z}}$, *denotada por la secuencia* $\{x_t\}_{t=1}^{T} = \{x_1, x_2, \ldots, x_T\}$.

Bajo esta concepción, cada observación x_t *se interpreta como un valor muestral o realización de la variable aleatoria* X_t *para un instante temporal concreto.*

Es fundamental señalar que, en la práctica, una serie temporal representa una única trayectoria parcial del proceso estocástico; es decir, para cada variable aleatoria X_t se dispone únicamente de una muestra de tamaño $n = 1$, lo que impide el uso de técnicas de inferencia convencionales basadas en la repetición del experimento en el mismo instante t. Asimismo, las observaciones de una serie temporal se caracterizan por estar dispuestas de forma ordenada y, habitualmente, espaciadas en intervalos cronológicos regulares (diarios, mensuales, trimestrales, etc.). El objetivo central del analista consiste en utilizar esta trayectoria observada para inferir las propiedades probabilísticas del proceso estocástico subyacente que la ha generado (Box et al., 2015; Hamilton, 1994)

La formulación algebraica de la dinámica temporal requiere la definición de los siguientes operadores lineales:

Definición A.3. *El **operador de retardo** B se define como un operador lineal que actúa sobre un elemento de una serie temporal desplazando el índice temporal una unidad hacia el pasado:* $Bx_t = x_{t-1}$. *Por extensión, para cualquier entero* k, *la aplicación sucesiva del operador resulta en* $B^k x_t = x_{t-k}$.

Definición A.4. *Se define el* ***operador diferencia regular de orden*** *d, denotado por ∇^d o $(1-B)^d$, como la aplicación sucesiva de d tasas de variación absolutas sobre los elementos de un proceso estocástico X_t, donde B representa el operador de retardo.*

El caso más común en el análisis de series temporales se da para un primer orden de diferenciación ($d = 1$), el cual se expresa algebraicamente como:

$$\nabla X_t = (1 - B)X_t = X_t - X_{t-1}.$$

Para órdenes superiores, la expresión se obtiene mediante el desarrollo del binomio de Newton para $(1 - B)^d$.

Definición A.5. *Se define el* ***operador diferencia estacional de orden*** *D* ***y período*** *s, denotado por ∇_s^D o $(1 - B^s)^D$, como la aplicación del operador diferencia sobre elementos de la serie temporal distanciados por un intervalo constante de s períodos.*

El caso más habitual en la práctica econométrica se presenta para un primer orden de diferenciación estacional ($D = 1$), el cual se expresa algebraicamente como:

$$\nabla_s X_t = (1 - B^s)X_t = X_t - X_{t-s}.$$

Por lo general, el parámetro s se fija en función de la frecuencia de muestreo de los datos, siendo $s = 12$ para series de periodicidad mensual y $s = 4$ para registros trimestrales.

A.2. Estructura de Momentos y Dependencia Lineal

Definición A.6. *Sea $\{X_t\}$ un proceso estocástico. Se definen sus* ***momentos*** *de primer y segundo orden como:*

- ***Función de medias*** *(μ_t): Representa la esperanza matemática de las distribuciones marginales del proceso para cada instante t:*

  $$\mu_t = \mathbb{E}[X_t],$$

  Un proceso se denomina ***estable en media*** *si esta función es constante ante desplazamientos temporales.*

- ***Función de varianzas*** *(σ_t^2): Proporciona la dispersión de cada variable del proceso respecto a su media:*

  $$\sigma_t^2 = Var(X_t) = \mathbb{E}[(X_t - \mu_t)^2].$$

  El proceso es ***estable en varianza*** *si σ_t^2 es constante para todo t.*

- ***Función de autocovarianzas*** *($\gamma_{s,t}$): Mide la dependencia lineal entre las variables del proceso en dos instantes cualesquiera s y t:*

  $$\gamma_{s,t} = Cov(X_s, X_t) = \mathbb{E}[(X_s - \mu_s)(X_t - \mu_t)].$$

  Nótese que, por definición, la varianza es un caso particular de esta función cuando $s = t$, es decir, $\gamma_{t,t} = \sigma_t^2$.

Definición A.7. *Para facilitar la interpretación de la memoria del proceso, se definen medidas estandarizadas de dependencia lineal:*

- ***Función de autocorrelación simple (ACF)**: Representa la correlación entre X_s y X_t y se obtiene normalizando la autocovarianza:*

$$Corr(X_s, X_t) = \rho_{s,t} = \frac{\gamma_{s,t}}{\sigma_s \sigma_t}.$$

Esta función ($\rho_{s,t} \in [-1, 1]$) permite medir la capacidad predictiva en el caso lineal de un valor pasado sobre uno futuro.

- ***Función de autocorrelación parcial (PACF)**: Mide la correlación entre X_s y X_t tras eliminar el efecto lineal ejercido sobre ellas por las variables intermedias $\{X_{s+1}, \ldots, X_{t-1}\}$. Se define formalmente como la correlación condicionada entre dichas variables:*

$$\alpha_{s,t} = Corr(X_s, X_t | X_{s+1}, \ldots, X_{t-1}),$$

o bien como:

$$\alpha(s,t) = \frac{Cov(X_s - \hat{X}_s, X_t - \hat{X}_t)}{\sqrt{Var(X_s - \hat{X}_s) Var(X_t - \hat{X}_t)}},$$

donde $\hat{X}_s$ y $\hat{X}_t$ representan los mejores predictores lineales de X_s y X_t, respectivamente, construidos a partir del conjunto de variables comprendidas entre los instantes s y t.

A.3. Estacionariedad, Ergodicidad y Teorema de Wold

La inferencia estadística sobre una única realización temporal exige imponer condiciones de regularidad que garanticen la invarianza temporal de la distribución paramétrica.

Definición A.8. *Un proceso estocástico $\{X_t\}$ es **estrictamente estacionario** si la distribución conjunta de cualquier colección de variables $\{X_{t_1}, X_{t_2}, \ldots, X_{t_n}\}$ es idéntica a la de la colección desplazada en el tiempo $\{X_{t_1+k}, X_{t_2+k}, \ldots, X_{t_n+k}\}$, para cualquier elección de instantes $t_1, ..., t_k$ y cualquier retardo k.*

La condición de **estacionariedad estricta** resulta, en la práctica, excesivamente restrictiva para la mayoría de las series económicas y naturales. Esta rigidez se debe a que implica que todas las variables $\{X_t\}$ deben estar **marginalmente idénticamente distribuidas**, lo que obliga a que no solo la media y la varianza sean constantes, sino también todos los momentos superiores y la probabilidad de eventos extremos en cada instante t. En entornos reales, factores como la evolución de los sistemas sociales (cambios de normativa o leyes), la estacionalidad marcada o los cambios bruscos en la volatilidad de los mercados invalidan frecuentemente esta premisa.

Definición A.9. *Un proceso estocástico $\{X_t\}$ se dice **débilmente estacionario, o estacionario de segundo orden**, si cumple tres requisitos:*

1. *La función de media es finita y constante : $\mathbb{E}[X_t] = \mu < \infty, \forall t$.*
2. *La función de varianza es finita y constante: $Var(X_t) = \sigma^2 < \infty, \forall t$.*
3. *La función de autocovarianza entre dos instantes depende únicamente del desfase o retardo k entre ellos, y no del instante concreto t: $Cov(X_t, X_{t+k}) = \gamma_k$.*

Para simplificar, en el presente trabajo se utilizará el término estacionario para referirnos en todo momento a la estacionariedad en sentido débil.

La **ergodicidad** es la propiedad que permite realizar inferencias sobre un proceso estocástico completo a partir de una única trayectoria observada. Matemáticamente, un proceso estacionario en covarianzas se define como ergódico si sus promedios temporales convergen a sus momentos poblacionales (esperanzas de conjunto). En esencia, la ergodicidad actúa como una generalización de la **Ley de los Grandes Números** para datos dependientes: esto justifica sustituir las esperanzas matemáticas inobservables por promedios calculados a lo largo del tiempo sobre una única realización $\{x_t\}_{t=1}^{T}$, asumiendo que dicha serie histórica es lo suficientemente informativa sobre la estructura global del proceso, $\{X_t\}$.

Teorema A.1 (Teorema Ergódico de von Neumann, 1932)**.** *Sea $\{X_n\}_{n\in\mathbb{Z}}$ un proceso estacionario tal que $X_n \in L^2$. Entonces, el promedio temporal converge en media cuadrática:*

$$\bar{X}_n = \frac{1}{n}\sum_{k=1}^{n} X_k \xrightarrow{L^2} \mathbb{E}[X_t|\mathcal{I}].$$

Donde $\mathcal{I}$ denota la σ-álgebra de eventos invariantes asociada al proceso estocástico. En el caso de que el proceso sea estrictamente ergódico, la estructura de $\mathcal{I}$ es trivial (contiene únicamente conjuntos de probabilidad cero o uno), lo que implica que la información histórica no altera el promedio a largo plazo. En consecuencia, la esperanza condicional colapsa en la media poblacional del proceso, cumpliéndose que:

$$\bar{X}_n \xrightarrow{L^2} \mu.$$

Teorema A.2 (Teorema Ergódico de Birkhoff, 1931)**.** *Sea $(\Omega, \mathcal{A}, P)$ un espacio de probabilidad y $T : \Omega \to \Omega$ una transformación del espacio sobre sí mismo que preserva la medida. Si $X \in L^1$, entonces, entonces el promedio temporal converge casi seguramente (con probabilidad 1) al la esperanza matemática:*

$$\lim_{n\to\infty} \frac{1}{n}\sum_{k=0}^{n-1} X(T^k(\omega)) = \mathbb{E}[X_t|\mathcal{I}] \quad c.s.$$

donde $\mathcal{I}$ es la σ-álgebra de los conjuntos invariantes bajo la transformación T. Para procesos ergódicos, este límite coincide con la esperanza matemática μ.

Definición A.10. *Un proceso estocástico $\{a_t\}$ se denomina **ruido blanco** (en sentido débil) si sus elementos tienen media cero, varianza constante y están serialmente incorrelacionados:*

1. $\mathbb{E}[a_t] = 0, \quad \forall t.$
2. $Var(a_t) = \mathbb{E}[a_t^2] = \sigma_a^2 < \infty, \quad \forall t.$
3. $\gamma_k = \mathbb{E}[a_t a_{t-k}] = 0 \quad$ *para todo* $k \neq 0.$

Definición A.11. *Sea $\{X_t\}_{t=1}^{T}$ una muestra de tamaño T proveniente de un proceso ergódico. Se definen los siguientes estimadores muestrales:*

- ***Media muestral ($\bar{X_T}$):** Es un estimador consistente de la media poblacional $\mu = \mathbb{E}[X_t]$:*

$$\bar{X_T} = \frac{1}{T}\sum_{t=1}^{T} X_t.$$

- ***Autocovarianza muestral ($\hat{\gamma}_k$):** Mide la covarianza observada en el retardo k. Siguiendo a Box et al. (2015), se utiliza T como divisor para garantizar que la matriz de autocovarianzas sea definida positiva:*

$$\hat{\gamma}_k = \frac{\sum_{t=1}^{T-k}(X_t - \bar{X}_T)(X_{t+k} - \bar{X}_T)}{T}, \quad k = 0, 1, \ldots, K.$$

- ***Autocorrelación simple muestral*** *($\hat{\rho}_k$): Se obtiene normalizando la autocovarianza por la varianza muestral ($\hat{\gamma}_0$):*

$$\hat{\rho}_k = \frac{\hat{\gamma}_k}{\hat{\gamma}_0}.$$

- ***Autocorrelación parcial muestral*** *($\hat{\alpha}_{kk}$): Se define como el último coeficiente estimado ($\hat{\alpha}_{kk}$) de una regresión lineal de X_t sobre sus k valores pasados:*

$$X_t = c + \hat{\alpha}_{k1}X_{t-1} + \hat{\alpha}_{k2}X_{t-2} + \cdots + \hat{\alpha}_{kk}X_{t-k} + \hat{e}_t.$$

Definición A.12. *Un* ***proceso estocástico*** $\{X_t\}_t$ ***es lineal*** *si puede ser representado como una combinación lineal ponderada de las observaciones de un proceso de ruido blanco, $\{a_t\}$. Es decir, si admite la siguiente representación:*

$$X_t = c + \sum_{i=-\infty}^{\infty} \psi_i a_{t-i} = \cdots + \psi_{-1}a_{t+1} + \psi_0 a_t + \psi_1 a_{t-1} + \ldots,$$

donde la secuencia de coeficientes $\{\psi_i\}$ debe ser ***absolutamente sumable****, cumpliendo que $\sum_{i=-\infty}^{\infty} |\psi_i| < \infty$.*

Definición A.13. *Un proceso estocástico $\{X_t\}_t$ es* ***causal (o*** *$MA(\infty)$) si puede representarse como:*

$$X_t = c + \psi_0 a_t + \psi_1 a_{t-1} + \psi_2 a_{t-2} \ldots, con \sum_{i=0}^{\infty} |\psi_i| < \infty.$$

Definición A.14. *Un proceso estocástico $\{X_t\}_t$ es* ***invertible (o*** *$AR(\infty)$) si puede representarse como:*

$$X_t = c + a_t + \pi_1 X_{t-1} + \pi_2 X_{t-2} + \ldots, con \sum_{i=1}^{\infty} |\pi_i| < \infty.$$

Teorema A.3 (Teorema de Descomposición de Wold)**.** *Formulado originalmente por Wold (1938), este teorema establece que cualquier proceso estocástico $\{X_t\}$ que sea estacionario en covarianza admite una representación única como la suma de dos componentes linealmente incorrelacionadas:*

$$X_t = \sum_{j=0}^{\infty} \psi_j a_{t-j} + \eta_t,$$

donde:

1. *$\{a_t\}$ es una secuencia de ruido blanco con varianza σ_a^2, que representa las* ***innovaciones*** *del proceso; es decir, el error cometido al realizar la mejor predicción lineal de X_t a partir de sus valores pasados.*
2. *Los coeficientes cumplen que $\psi_0 = 1$ y son cuadrado-sumables ($\sum_{j=0}^{\infty} \psi_j^2 < \infty$).*
3. *η_t es la componente* ***linealmente determinista****, definida como aquella parte del proceso que puede predecirse con error nulo a partir de su pasado histórico (por ejemplo, constantes, tendencias o funciones senoidales).*

Si el proceso carece de componente determinista ($\eta_t = 0$), se denomina ***no determinista lineal puro****. En este caso, el proceso admite una representación de medias móviles de orden infinito, $MA(\infty)$:*

$$X_t = a_t + \psi_1 a_{t-1} + \psi_2 a_{t-2} + \ldots.$$

A.4. Modelos Box-Jenkins

Una vez introducidas las series de tiempo y sus características y particularidades, se procede con la evolución teórica lógica hacia distintos procesos estocásticos como posibles generadores de las mismas, en concreto los modelos de la metodología Box-Jenkins (Box et al., 2015) ya nombrados anteriormente.

A.4.1. Modelos para series estacionarias

Comenzamos estudiando los modelos para series estacionarias apoyándose en el tercer capítulo de Box et al. (2015, pp. 47–88).

Definición A.15. *En un proceso* ***autorregresivo de orden*** p***, AR(p)****, el valor de la serie en el instante t se explica mediante una combinación lineal de sus propios valores pasados hasta el retardo p, más un término de error o innovación, generalizando los modelos de regresión lineal:*

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

donde c,$\phi_1, \ldots, \phi_p$ ($\phi_p \neq 0$) y $\{a_t\}_t$ son parámetros de un proceso de ruido blanco.

Definición A.16. *Dado un proceso AR(p) se definen:*

- ***Operador autorregresivo****: $\phi(z) = 1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p$, donde B es el operador retardo.*
- ***Polinomio AR(p)*** *característico: $\phi(z) = 1 - \phi_1 z - \phi_2 z^2 - \cdots - \phi_p z^p$. A partir de la definición del operador autorregresivo se puede escribir la representación del proceso AR en su forma compacta: $\phi(B)X_t = c + a_t$*

Para que el proceso AR(p) sea considerado estacionario las raíces del polinomio característico $\Phi_p(z) = 1 - \phi_1 z - \cdots - \phi_p z^p = 0$ deben tener módulo distinto de uno. Además, el proceso será causal si las raíces de este mismo polinomio característico tienen un módulo mayor que uno. Por último, de acuerdo con la Definición A.14, un proceso AR(p) es siempre invertible sin más que tomar $\pi_i = \phi_i$ con i=1,...,p y nulos los demás hasta infinito.

Definición A.17. *En un proceso de* ***medias móviles de orden*** q***, MA(q)****, la observación actual se define como una combinación lineal de la innovación presente y de las q innovaciones pasadas:*

$$X_t = c + a_t + \theta_1 a_{t-1} + \cdots + \theta_q a_{t-q},$$

donde c representa el término constante del proceso, $\theta_1, \theta_2, \ldots, \theta_q$ son constantes ($\theta_q \neq 0$) y $\{a_t\}$ es un proceso de ruido blanco con varianza σ_a^2.

Definición A.18. *Dado un proceso MA(q) se definen:*

- ***Operador de medias móviles****: $\theta(B) = 1 + \theta_1 B + \theta_2 B^2 + \cdots + \theta_p B^q$, donde B es el operador retardo.*
- ***Polinomio MA(q)*** *característico: $\theta(z) = 1 + \theta_1 z + \theta_2 z^2 + \cdots + \theta_p z^q$.*

  A partir de la definición del operador autorregresivo se puede escribir la representación del proceso AR en su forma compacta: $X_t = c + \theta(B)a_t$

Por construcción, cualquier proceso $MA(q)$ finito es siempre lineal y causal, siguiendo las definiciones (A.12) y (A.13) y tomando $a_i = \theta_i$ con $i = 1, \ldots, q$ y nulos los restantes. Adicionalmente, una condición suficiente para que este proceso sea invertible es que las raíces del polinomio característico $MA(q)$ estén fuera del círculo unitario.

Definición A.19. *Un **proceso estacionario** $\{X_t\}_t$ **puede representarse como un modelo ARMA de órdenes p y q, ARMA(p,q)**, combinando las estructuras autorregresiva y de medias móviles y admitiendo la siguiente representación:*

$$X_t = c + \phi_1 X_{t-1} + \phi_2 X_{t-2} + \cdots + \phi_p X_{t-p} + a_t + \theta_1 a_{t-1} + \theta_2 a_{t-2} + \cdots + \theta_q a_{t-q}, \tag{A.1}$$

siendo $\{X_t\}_t$ un proceso estocástico estacionario, $c, \phi_1, \phi_2, \ldots, \phi_p, \theta_1, \theta_2, \ldots, \theta_q$ constantes con $\phi_p \neq 0, \theta_q \neq 0$ y $\{a_t\}_t$ un proceso de ruido blanco con varianza σ_a^2.

Se puede observar que empleando los operadores autorregresivos y de medias móviles la ecuación (A.1) del modelo puede escribirse de forma compacta como: $\phi(B)X_t = c + \theta(B)a_t$

La caracterización de un proceso $ARMA(p, q)$ exige el cumplimiento de requisitos estructurales que garantizan su estabilidad y representatividad. La condición de estacionariedad es equivalente a que las raíces del polinomio autorregresivo $\phi(z) = 1 - \phi_1 z - \phi_2 z^2 - \cdots - \phi_p z^p$ no tengan módulo unidad ($\phi(z) \neq 0, \forall |z| = 1$); sin embargo, para asegurar la causalidad del sistema -esto es, que el proceso sea representable como un $MA(\infty)$ de innovaciones pasadas- se requiere que dichas raíces se sitúen estrictamente fuera del círculo unidad ($\phi(z) \neq 0 \forall |z| \leq 1$). De manera análoga, el proceso será invertible $\iff$ las raíces del polinomio de medias móviles $\theta(z) = 1 + \theta_1 z + \theta_2 z^2 + \cdots + \theta_q z^q$ se sitúan fuera del círculo unitario ($\theta(z) \neq 0, \forall |z| \leq 1$), permitiendo expresar la innovación actual como una combinación lineal de los valores observados. Bajo estas premisas, el término constante verifica $c = \mu(1 - \phi_1 - \cdots - \phi_p)$, donde μ es la media del proceso. Finalmente, dado que las funciones de autocorrelación simple (ACF) y parcial (PACF) de un $ARMA(p, q)$ son el resultado de superponer las dinámicas de sus componentes base, su identificación visual a partir de los correlogramas se torna una tarea significativamente compleja, reduciéndose a patrones conocidos únicamente en los casos particulares donde p=0 y/o q=0.

Definición A.20. *Un **modelo ARMA estacional** de órdenes P y Q con período s, denotado como $ARMA(P, Q)_s$, se define mediante la ecuación:*

$$X_t = c + \Phi_1 X_{t-s} + \Phi_2 X_{t-2s} + \cdots + \Phi_P X_{t-Ps} + a_t + \Theta_1 a_{t-s} + \cdots + \Theta_Q a_{t-Qs}, \tag{A.2}$$

*donde $\{X_t\}_t$ es un proceso estocástico **débilmente estacionario**, c es una constante, y los parámetros estacionales $\{\Phi_1, \ldots, \Phi_P\}$ y $\{\Theta_1, \ldots, \Theta_Q\}$ cumplen estrictamente que $\Phi_P \neq 0$ y $\Theta_Q \neq 0$, siendo $\{a_t\}_t$ un proceso de ruido blanco.*

Es fundamental observar que la ecuación (A.2) puede expresarse de forma compacta mediante los operadores estacionales presentados en la definición (A.5):

$$\Phi_P(B^s)X_t = c + \Theta_Q(B^s)a_t,$$

donde los polinomios estacionales vienen dados por $\Phi_P(B^s) = 1 - \Phi_1 B^s - \cdots - \Phi_P B^{Ps}$ y $\Theta_Q(B^s) = 1 + \Theta_1 B^s + \cdots + \Theta_Q B^{Qs}$. Desde un punto de vista analítico, este modelo equivale a un $ARMA$ convencional de órdenes muy elevados donde los coeficientes correspondientes a retardos no múltiplos de s están restringidos a cero, heredando así todas las propiedades de estabilidad e invertibilidad discutidas anteriormente.

Definición A.21. *Un **modelo ARMA estacional multiplicativo**, denotado como $ARMA(p, q) \times (P, Q)_s$, permite capturar la interacción entre la componente regular y la estacional mediante el producto de sus operadores:*

$$\Phi_P(B^s)\phi_p(B)X_t = c + \Theta_Q(B^s)\theta_q(B)a_t,$$

donde $\phi_p(B)$ y $\theta_q(B)$ representan la dinámica regular, mientras que $\Phi_P(B^s)$ y $\Theta_Q(B^s)$ modelizan la estructura estacional.

A.4.2. Modelos para series no estacionarias

En los apartados previos se han analizado diversos procesos candidatos a actuar como generadores de series estacionarias. No obstante, en la práctica empírica, y especialmente en el análisis macroeconómico, es infrecuente hallar series que satisfagan de origen las hipótesis de estacionariedad descritas. Los factores determinantes del incumplimiento de dichas premisas suelen clasificarse en tres categorías:

1. **Heterocedasticidad**: La varianza del proceso no permanece constante a lo largo del tiempo, violando la condición $\mathbb{V}\text{ar}(X_t) = \sigma^2 < \infty$ para todo t.
2. **Tendencia**: El primer momento poblacional presenta dependencia temporal, es decir, $\mathbb{E}[X_t] = \mu_t$, indicando no estacionariedad en la media.
3. **Componente estacional**: La esperanza matemática del proceso fluctúa de manera periódica en función de una pauta cíclica regular, tal que $\mathbb{E}[X_t - tendencia] = \mu_{t\pm s}$ para un periodo estacional s.

A través del examen del gráfico secuencial de la serie[1], es posible diagnosticar la **inestabilidad en la varianza**. Este concepto suele manifestarse cuando la varianza del proceso es proporcional a su nivel (comportamiento de tipo multiplicativo). **Bajo estas condiciones específicas de heterocedasticidad dependiente del nivel**, la serie puede ser estabilizada mediante la familia de transformaciones propuestas por Box y Cox (1964), la cual se define formalmente como:

$$X_t^{(\lambda)} = \begin{cases} \frac{X_t^{\lambda}-1}{\lambda} & \text{si } \lambda \neq 0 \\ \ln(X_t) & \text{si } \lambda = 0, \end{cases}$$

donde λ es el parámetro de transformación.

Otro motivo por el que una serie puede no ser estacionaria se debe a la presencia de una **tendencia sistemática en su nivel**, lo que implica una dependencia directa entre la media del proceso y el transcurso del tiempo. Para corregir este fenómeno, se recurre a la técnica de la diferenciación regular de la serie.

Definición A.22. *Un proceso estocástico $\{X_t\}_t$ no estacionario en media se dice que es **integrado de orden** d ($d > 0$), denotado como $\{X_t\}_t \sim I(d)$, si al aplicar el operador diferencia regular de orden d se obtiene un proceso estacionario, mientras que con un orden $d-1$ el proceso persiste en su no estacionariedad. Matemáticamente:*

$$\nabla^d X_t = (1-B)^d X_t \sim I(0)$$

Bajo este marco conceptual se definen los modelos no estacionarios más importantes, **los procesos integrados**, que tienen la propiedad fundamental que al diferenciarlos se obtienen procesos estacionarios.

Definición A.23. *Un modelo **autorregresivo integrado de medias móviles**, $ARIMA(p,d,q)$, se expresa mediante la interacción de los operadores autorregresivos y de medias móviles sobre la serie diferenciada d veces:*

$$\phi_p(B)(1-B)^d X_t = c + \theta_q(B)a_t,$$

donde c es una constante, $\phi_p(B)$ y $\theta_q(B)$ son los polinomios característicos definidos anteriormente y $\{a_t\}_t$ es un proceso de ruido blanco.

[1] Representación de las observaciones de una serie x_t ordenadas cronológicamente en el eje de abscisas, facilitando la identificación visual de la evolución dinámica y posibles cambios estructurales.

Es importante señalar que, en la práctica econométrica habitual, un orden de integración $d \leq 2$ suele ser suficiente para estabilizar la media de la serie. En la práctica, superar erróneamente este orden podría inducir un fenómeno de sobrediferenciación, eliminando información relevante de la señal económica.

Por último, de forma análoga a la tendencia regular, la **estacionalidad** puede inducir no estacionariedad si el nivel medio de la serie varía en función del periodo estacional s. En este contexto, surge la necesidad de la diferenciación estacional.

Definición A.24. *Un proceso $\{X_t\}_t$ se define como **estacionalmente integrado de orden** D ($D > 0$) si requiere la aplicación del operador diferencia estacional $(1-B^s)^D$ para alcanzar la estacionariedad. Los modelos que incorporan esta estructura se denominan ARIMA estacionales:*

$$\Phi_P(B^s)(1-B^s)^D X_t = c + \Theta_Q(B^s)a_t.$$

La síntesis de todas las estructuras anteriores da lugar a los modelos ARIMA estacionales multiplicativos. Estos modelos son los más versátiles, ya que engloban todos los casos previos como situaciones particulares y son capaces de capturar simultáneamente la memoria de corto plazo y los ciclos estacionales.

Definición A.25. *Un modelo **ARIMA estacional multiplicativo, denotado como** $ARIMA(p,d,q)\times(P,D,Q)_s$, se formula combinando todos los operadores de diferenciación y polinomios de retardo:*

$$\Phi_P(B^s)\phi_p(B)(1-B^s)^D(1-B)^d X_t = c + \Theta_Q(B^s)\theta_q(B)a_t.$$

A.5. Identificación, Selección y Predicción

La identificación de la estructura paramétrica requiere una secuencia de diagnosis rigurosa.

A.5.1. Contrastes de Estacionariedad (Raíz Unitaria)

Para validar el orden de integración d, se emplean contrastes estadísticos formales:

- **Test de Dickey-Fuller Aumentado (ADF)**;(Said y Dickey, 1984): Contrasta la hipótesis nula de existencia de raíz unitaria ($H_0 : \gamma = 0$). Donde γ es el coeficiente del primer retardo de la serie sin diferenciar. Si se rechaza, la serie es estacionaria.
- **Test KPSS** (Kwiatkowski, Phillips, Schmidt, y Shin, 1992): A diferencia del Test ADF, su hipótesis nula es la estacionariedad ($H_0 : X_t \sim I(0)$). El uso conjunto de ambos tests permite confirmar si una serie es realmente estacionaria o si requiere una diferencia adicional.

A.5.2. Distribuciones Asintóticas y Criterios de Selección

La identificación del orden del proceso generador se basa en el comportamiento de las funciones de autocorrelación simple (ACF) y parcial (PACF). No obstante, en el análisis empírico no se dispone de los coeficientes poblacionales, por lo que la identificación debe realizarse a partir de sus estimadores muestrales. Según Cryer y Chan (2008), bajo el supuesto de que el tamaño muestral T es suficientemente elevado, las distribuciones de estos estimadores permiten definir bandas de confianza estadísticas para determinar el **punto de truncamiento de las funciones de autocorrelación**:

- **Procesos AR(p)**: Los coeficientes de la PACF para retardos superiores a p se distribuyen asintóticamente como:

$$\hat{\alpha}_k \sim N\left(0, \frac{1}{T}\right), \quad \forall k > p.$$

- **Procesos MA(q)**: Los coeficientes de la ACF para retardos superiores a q siguen una distribución:

$$\hat{\rho}_k \sim N\left(0, \frac{1 + 2\sum_{i=1}^{q} \rho_i^2}{T}\right), \quad \forall k > q.$$

Frecuentemente, la etapa de identificación sugiere múltiples modelos válidos. Para decidir entre ellos, se aplica el principio de **parsimonia** (Peña, 2010): el modelo más sencillo es preferible para evitar el sobreajuste (*overfitting*). Para cuantificar este equilibrio, se utilizan los Criterios de Información, que penalizan la verosimilitud en función del número de parámetros k:

- **Criterio de Información de Akaike (AIC)** (Akaike, 1974): se enfoca en minimizar la pérdida de información y es especialmente útil en contextos de predicción o *nowcasting*, aunque tiende a seleccionar modelos con un número de parámetros ligeramente superior al óptimo.

$$AIC = -2\ln(\hat{L}) + 2k.$$

- **Criterio de Información de Akaike Corregido (AICc)** (Sugiura, 1978): Es una versión del AIC que incluye una corrección para muestras finitas, diseñada para reducir el sesgo cuando el tamaño muestral T es pequeño en relación al número de parámetros k. A medida que T aumenta, el AIC_c converge asintóticamente al AIC.

$$AIC_c = -2\ln(\hat{L}) + \frac{2(kT + k + 2)}{T - k - 2}.$$

- **Criterio Bayesiano de Schwarz (BIC)** (Schwarz, 1978): penaliza más fuertemente la complejidad en función del tamaño muestral T. Es un criterio consistente que tiende a seleccionar modelos más compactos, evitando el sobreajuste en paneles de alta dimensionalidad.

$$BIC = -2\ln(\hat{L}) + k\ln(T),$$

donde $\hat{L}$ representa el máximo de la función de verosimilitud.

A.5.3. Contrastes de Diagnosis Residual

La validez de la inferencia estadística tradicional –en particular, la precisión de las bandas de los intervalos de predicción y los contrastes de significatividad en muestras finitas– depende de que los residuos se comporten como un ruido blanco gaussiano (A.10). Es importante destacar que la ausencia de normalidad no invalida la consistencia de los estimadores del modelo ni la optimalidad de las predicciones puntuales, las cuales se sostienen bajo la teoría de Cuasi-Máxima Verosimilitud (QMLE), pero sí exige cautela al interpretar los intervalos probabilísticos asintóticos. Se verifican mediante los siguientes contrastes:

1. **Incorrelación (Test de Ljung-Box)**; Ljung y Box (1978): Analiza si las primeras h autocorrelaciones de los residuos son conjuntamente nulas. Bajo la hipótesis nula de ausencia de autocorrelación, el estadístico Q se distribuye como una chi-cuadrado con $h - p - q$ grados de libertad:

$$Q = T(T+2)\sum_{k=1}^{h} \frac{\hat{\rho}_k^2(\hat{a})}{T-k} \sim \chi^2_{h-(p+q)},$$

donde p y q son los órdenes de las partes autorregresiva y de medias móviles

2. **Contraste de Media Nula**: Verifica si el valor esperado de los residuos es significativamente distinto de cero mediante el estadístico $t = \bar{a}/(\hat{\sigma}_a/\sqrt{T})$. Bajo $H_0 : \mathbb{E}[a_t] = 0$, este estadístico sigue una distribución t-Student asintóticamente normalizada.

3. **Normalidad**: Verificada mediante el test de **Jarque-Bera** (Jarque y Bera, 1987) y el contraste de **Shapiro-Wilk** (Shapiro y Wilk, 1965). La normalidad es fundamental para la construcción de los intervalos de predicción.

4. **Homocedasticidad**: Se comprueba que la varianza de los residuos sea constante. La detección de estructuras no lineales o cambios en la volatilidad es fundamental para descubrir si existen patrones no lineales dentro de los datos con los que se trabajan.

A.5.4. Predicción y Precisión

El objetivo último es proyectar el comportamiento futuro de la serie. El predictor óptimo que minimiza el error cuadrático medio (MSE) -tanto para la secuencia observada como para el promedio de todas las posibles realizaciones del proceso estocástico- es la esperanza condicionada al conjunto de información disponible $\mathcal{I}_T$. Este resultado de optimización global se fundamenta en (Peña, 2010, pp. 224–226):

$$\hat{x}_{T+h|T} = \mathbb{E}[x_{T+h}|\mathcal{I}_T].$$

Para evaluar la precisión fuera de la muestra y permitir la comparación entre los modelos ARIMA y los modelos factoriales que se desarrollarán más adelante, se emplea la Raíz del Error Cuadrático Medio (RMSE):

$$RMSE = \sqrt{\frac{1}{h}\sum_{t=T+1}^{T+h}(x_t - \hat{x}_{t|T})^2}.$$

A.6. Extensión Multivariante: El Modelo VAR(p)

El análisis de sistemas multivariantes se fundamenta en la extensión del caso univariante al vectorial:

Definición A.26. *Se define una* ***serie de tiempo múltiple*** *como una sucesión de vectores ordenados secuecialmente de i variables aleatorias observadas en el instante t:*

$$\mathbf{y}_t' = \Big(y_{1t},\ y_{2t},\ \dots\ y_{it}\Big) \in \mathbb{R}^i, \quad t \in \mathcal{I},$$

donde cada componente y_{it} representa una realización de un proceso estocástico univariante. Siendo i la dimensión del sistema, es decir, el número total de series temporales que se analizan de forma simultánea.

Definición A.27. *Un proceso vectorial $\{\mathbf{y}_t\}$ es* ***débilmente estacionario*** *si su vector de medias es constante, $\mathbb{E}[\mathbf{y}_t] = \boldsymbol{\mu}$, y su matriz de autocovarianzas depende únicamente del retardo l:*

$$\boldsymbol{\Gamma}(l) = \mathbb{E}[(\mathbf{y}_t - \boldsymbol{\mu})(\mathbf{y}_{t-l} - \boldsymbol{\mu})'], \quad \forall t \in \mathcal{I},$$

donde $\boldsymbol{\Gamma}(0) = \boldsymbol{\Sigma}_y$ representa la matriz de varianzas-covarianzas contemporánea, es decir, en el instante actual.

Definición A.28. *Se define* ***un proceso de ruido blanco K-dimensional*** *$,\mathbf{u}_t$, como una secuencia de vectores aleatorios tales que $\mathbb{E}[\mathbf{u}_t] = \mathbf{0}$, $\mathbb{E}[\mathbf{u}_t\mathbf{u}_t'] = \boldsymbol{\Sigma}_u$ (siendo $\boldsymbol{\Sigma}_u$ una matriz definida positiva) y $\mathbb{E}[\mathbf{u}_t\mathbf{u}_s'] = \mathbf{0}$ para todo $t \neq s$.*

A.6.1. El Modelo de Vectores Autorregresivos (VAR)

El modelo de Vectores Autorregresivos constituye el pilar del análisis macroeconométrico multivariante desde la propuesta de (Sims, 1980). A diferencia de los modelos univariantes, el enfoque VAR captura la interdependencia dinámica entre múltiples series tratando a todas las variables del sistema como endógenas.

Definición A.29. *Sea* $\mathbf{y}_t$ *un vector de* K *variables endógenas, un modelo VAR(p) se define como:*

$$\mathbf{y}_t = \mathbf{c} + \mathbf{A}_1\mathbf{y}_{t-1} + \cdots + \mathbf{A}_p\mathbf{y}_{t-p} + \mathbf{u}_t,$$

donde $\mathbf{c}$ *es un vector fijo de constantes de dimensión* $K \times 1$ *(el intercepto del sistema),* $\mathbf{A}_i$ *son las matrices de coeficientes estructurales de tamaño* $K \times K$ *que capturan las relaciones contemporáneas y cruzadas, y* $\mathbf{u}_t$ *representa el vector de ruido blanco multivariante con matriz de varianzas-covarianzas* $\mathbf{\Sigma}_u$*.*

La validez empírica del sistema depende de su estabilidad. Formalmente, el proceso es **estable** si y solo si el polinomio característico no tiene ninguna raíz dentro o sobre el círculo unidad complejo, esto es:

$$\det(\mathbf{I}_K - \mathbf{A}_1 z - \cdots - \mathbf{A}_p z^p) \neq 0, \quad \forall |z| \leq 1.$$

En la práctica, el cumplimiento de esta condición garantiza que el impacto de los shocks aleatorios sea transitorio (disipándose asintóticamente hacia el equilibrio del sistema) y que el proceso vectorial sea estacionario en covarianzas (Lütkepohl, 2005).

Apéndice B

Notación de Álgebra Lineal y Matricial

Siguiendo la notación estándar de la literatura y a modo de guía se utilizan los siguientes símbolos y operadores a lo largo de esta memoria:

- **Vectores y Matrices**: Las letras en negrita minúscula ($\mathbf{y}, \mathbf{x}$) denotan vectores columna, mientras que las letras en negrita mayúscula ($\mathbf{A}, \mathbf{B}$) denotan matrices.
- **Transposición**: $\mathbf{A}'$ denota la transpuesta de la matriz $\mathbf{A}$.
- **Matriz Identidad** ($\mathbf{I}_K$): Matriz cuadrada de dimensión $K \times K$ con valores igual a uno en su diagonal principal y ceros en el resto de sus elementos.
- **Determinante y Traza**: $\det(\mathbf{A})$ representa el determinante de la matriz cuadrada $\mathbf{A}$, y $\text{tr}(\mathbf{A}) = \sum_{i=1}^{K} a_{ii}$ representa la traza de la misma (la suma de los elementos de su diagonal principal).
- **Producto de Kronecker** ($\mathbf{A} \otimes \mathbf{B}$): Operador que denota la multiplicación de matrices bloque a bloque.
- **Producto de Hadamard** ($\mathbf{A} \odot \mathbf{B}$): Operador que representa el producto elemento a elemento de dos matrices de igual dimensión.
- **Vectorización** ($\text{vec}(\mathbf{A})$): Operador lineal que transforma una matriz de dimensión $m \times n$ en un vector columna de dimensión $mn \times 1$ mediante el apilamiento secuencial de sus columnas.
- **Half-vectorization** ($\text{vech}(\mathbf{A})$): Operador que se aplica exclusivamente a matrices simétricas de dimensión $n \times n$ para apilar en un vector de dimensión $\left[\frac{n(n+1)}{2}\right] \times 1$ únicamente los elementos situados sobre la diagonal principal y por debajo de ella, eliminando así la redundancia paramétrica del sistema.
- **Descomposición de Cholesky**: Para una matriz $\mathbf{\Sigma}$ simétrica y definida positiva, existe una única matriz triangular inferior $\mathbf{L}$ con elementos diagonales estrictamente positivos tal que cumple la igualdad $\mathbf{\Sigma} = \mathbf{L}\mathbf{L}'$.

Apéndice C

Representación formal en el Espacio de Estados del Modelo M04

El Modelo Factorial Dinámico (MFD) de frecuencia mixta estimado para el *nowcasting* del PIB español se ejecuta sobre una muestra efectiva de $T = 376$ observaciones mensuales (desde febrero de 1995 hasta abril de 2026). Al considerar $N = 11$ series en el panel y un único factor común ($r = 1$) con $p = 2$ retardos en su ecuación de transición, el modelo se expresa en su forma estructural compacta como:

$$\mathbf{X}_t = \mathbf{\Lambda} F_t + \mathbf{e}_t.$$

$$F_t = A_1 F_{t-1} + A_2 F_{t-2} + u_t, \quad u_t \sim N(0, Q).$$

$$\mathbf{e}_t = \rho \mathbf{e}_{t-1} + \mathbf{v}_t, \quad \mathbf{v}_t \sim N(\mathbf{0}, \mathbf{\Sigma}_v).$$

Donde $\mathbf{X}_t, \mathbf{e}_t, \mathbf{v}_t \in \mathbb{R}^{11\times 1}$, la matriz de cargas factoriales $\mathbf{\Lambda} \in \mathbb{M}_{11\times 1}$, y las matrices de persistencia y varianzas de las innovaciones $\rho, \mathbf{\Sigma}_v \in \mathbb{M}_{11\times 11}$ tienen una estructura estrictamente diagonal. El factor común $F_t \in \mathbb{R}$, mientras que los parámetros dinámicos de transición A_1, A_2 y la varianza estructural Q son escalares. El conjunto de indicadores analizados se encuentra indexado por $i \in \{1, 2, \ldots, 11\}$, cuyas posiciones corresponden a las variables detalladas en el Cuadro C.1.

Estimación de los parámetros del sistema

Los valores paramétricos estimados mediante el algoritmo Esperanza-Maximización (EM) para los componentes de la ecuación de medida se recogen en el siguiente cuadro:

Por su parte, sustituyendo los coeficientes estimados para la dinámica del factor común autorregresivo, la ecuación de transición empírica que gobierna el ciclo económico latente queda determinada por:

$$F_t = 0{,}8398 F_{t-1} - 0{,}0571 F_{t-2} + u_t, \quad \text{con} \quad Var(u_t) = Q = 0{,}4382.$$

Para iniciar las recursiones del Filtro de Kalman en $t = 1$, el algoritmo inicializa los componentes a partir de la media y varianza incondicional estable del sistema en $t = 0$, dadas por los escalares óptimos:

$$\hat{F}_{0|0} = \mathbb{E}[F_0] = -0{,}9351, \qquad P_{0|0} = Var(F_0) = 0{,}2709.$$

Tabla C.1: Parámetros estimados de la Ecuación de Medida (Modelo M04)
.

Indicador ($\mathbf{X}_{it}$)	**Carga Factorial (λ_i)**	**Var. Innovación Aleatoria ($\sigma^2_{v,i}$)**	**Persistencia AR(1) (ρ_{ii})**
1. AFIL_SS	0.2786	0.7793	0.0150
2. VGE_TOT	0.3779	0.6755	-0.4160
3. IPI	0.3580	0.6810	-0.4311
4. IMP_BIENES	0.2624	0.8586	-0.1476
5. EXP_BIENES	0.2553	0.8164	-0.3736
6. ICN_SERV	0.3230	0.7509	-0.2963
7. CEM	0.2703	0.8330	-0.2531
8. VGE_SAL	0.3091	0.7852	-0.1687
9. PMI_MANU	0.3052	0.0769	0.9367
10. OCUPADOS	0.0428	0.0464	-0.0746
11. PIB	0.0615	0.0201	0.2157

Representación en el Espacio de Estados completa

Para resolver simultáneamente la presencia de frecuencias mixtas y la inercia autorregresiva de los residuos, la librería `dfms` expande el sistema hacia la representación homogénea de Bańbura y Modugno (2014). El vector de estado ampliado de este modelo de factor único posee una dimensión exacta de 24×1, configurado estructuralmente de la siguiente forma:

$$\boldsymbol{\alpha}_t = \begin{pmatrix} F_t & F_{t-1} & \dots & F_{t-4} & e_{1,t} & \dots & e_{9,t} & e_{10,t} & \dots & e_{10,t-4} & e_{11,t} & \dots & e_{11,t-4}. \end{pmatrix}'$$

La longitud de 24 elementos responde a la suma de los 5 espacios requeridos para la memoria del factor mensual, los 9 componentes de error autorregresivos de las series mensuales y los $5 \times 2 = 10$ espacios necesarios para almacenar los promedios móviles de las perturbaciones aleatorias de las variables trimestrales (`OCUPADOS` y `PIB`).

Sustituyendo los parámetros estimados del Cuadro C.1, la **Matriz de Observación Empírica $\mathbf{C}$** (11×24) del filtro queda definida numéricamente por bloques. Para las filas correspondientes a los 9 indicadores mensuales ($i = 1, \dots, 9$), la carga factorial es estrictamente contemporánea sobre la primera columna, mientras que el error sectorial activa una identidad en su respectiva posición:

$$\mathbf{C}_{i,\cdot} = \begin{pmatrix} \lambda_i & 0 & 0 & 0 & 0 & \dots & 1_{(col=5+i)} & \dots & 0. \end{pmatrix}$$

Por el contrario, para las variables trimestrales en log-diferencias ($i = 10$ para `OCUPADOS` e $i = 11$ para el `PIB`), la matriz distribuye los pesos de la agregación temporal de Mariano-Murasawa $(1/3, 2/3, 1, 2/3, 1/3)$ ponderados por sus respectivas cargas factoriales brutas, replicando idéntica estructura sobre los bloques de retardo de sus propios errores aleatorios. De este modo, la última fila de la matriz $\mathbf{C}$, correspondiente a la ecuación de medida del PIB, queda determinada por:

$$\begin{aligned} \text{PIB}_t =& 0{,}0615F_t + 0{,}1229F_{t-1} + 0{,}1843F_{t-2} + 0{,}1229F_{t-3} + 0{,}0615F_{t-4} \\ &+ \frac{1}{3}e_{11,t} + \frac{2}{3}e_{11,t-1} + e_{11,t-2} + \frac{2}{3}e_{11,t-3} + \frac{1}{3}e_{11,t-4}. \end{aligned}$$

Finalmente, de acuerdo con la parametrización de regularización numérica del algoritmo Quasi-EM, la matriz de covarianza de la ecuación de medida se fija de forma determinista como una matriz identidad

escalar de ruido infinitesimal, anulando la varianza de observación contemporánea en favor del bloque de transición:

$$\mathbf{R}_{fullstate} = 10^{-4} \cdot \mathbf{I}_{11}.$$

Las varianzas estimadas de la tercera columna del Cuadro C.1 se inyectan directamente sobre la diagonal principal de la matriz de covarianzas de la ecuación de transición $\mathbf{Q}_{fullstate}$ (24×24), en las posiciones correspondientes a las innovaciones estructurales puras $\mathbf{v}_t$, garantizando que toda la incertidumbre dinámica del panel sea gestionada de forma endógena por las ecuaciones predictivas del Filtro de Kalman.

Validación del supuesto de modelo aproximado

De acuerdo con lo formalizado por Chamberlain y Rothschild (1983) y extendido por Stock y Watson (2002b), la consistencia del estimador bajo especificaciones aproximadas exige verificar empíricamente que las perturbaciones compartan únicamente dependencias cruzadas de carácter débil. A partir de la matriz de varianzas-covarianzas de los residuos calculada en la estimación del modelo ($\hat{\mathbf{\Sigma}}_e$), se comprueban los dos requisitos necesarios del espacio espectral ya introducidos en la Sección 3.3.2.3:

1. *Acotación geométrica del ruido:* Al resolver la ecuación del operador lineal $|\hat{\mathbf{\Sigma}}_e - \lambda \mathbf{I}| = 0$, se extrae el espectro ordenado de autovalores del modelo: El autovalor máximo es $\lambda_{\text{máx}}(\hat{\mathbf{\Sigma}}_e) = 1{,}7691$. Al

Tabla C.2: Espectro ordenado de autovalores de la matriz de residuos ($\hat{\mathbf{\Sigma}}_e$)

.

Componente (i)	**Autovalor** (λ_i)
1	1,7691
2	1,1044
3	0,9037
4	0,7528
5	0,6080
6	0,5110
7	0,4913
8	0,3571
9	0,3285
10	0,0102
11	0,0072

encontrarse acotado muy cerca de la unidad, se valida la ausencia de una segunda componente común dominante no extraída por el sistema, validando la selección del factor único que se indicaba en la Sección 3.3.4

2. *Sumabilidad absoluta de la sección transversal:* Esta condición exige que las relaciones entre los errores de las distintas variables sean débiles y locales. En la práctica, si sumamos la relación (covarianza en valor absoluto) del error de una variable frente a todas las demás, el resultado no debe dispararse, sino mantenerse bajo un límite. Esto garantiza que, al promediar los indicadores para extraer el factor común, los ruidos individuales se cancelen entre sí en lugar de acumularse. Operando sobre los elementos de la sección cruzada mediante la expresión $\sum_{j \neq i} |Cov(e_{it}, e_{jt})|$, se presentan los cálculo en la Tabla C.3:

Los resultados demuestran que las dependencias mutuas remanentes se concentran exclusivamente en el clúster comercial e industrial, manteniéndose numéricamente acotadas bajo la frontera técnica del sistema. Ambas verificaciones ratifican que el panel se comporta formalmente como un Modelo de Factores Dinámicos Aproximado. Cabe destacar la consistencia que guarda este espectro residual frente a la descomposición de los datos originales presentada en la Tabla 5.5. Al extraer un único factor

Tabla C.3: Sumabilidad absoluta de covarianzas residuales fuera de la diagonal

.

Variable (i)	**Suma Acumulada Absoluta** ($\sum \lvert Cov(e_i, e_j) \rvert$)
Clúster de mayor interdependencia	
(IPI)	1,3336
(VGE_TOT)	1,2665
(EXP_BIENES)	1,2520
(ICN_SERV)	1,0146
(CEM)	0,9848
Clúster de interdependencia moderada	
(IMP_BIENES)	0,6538
(VGE_SAL)	0,6236
(AFIL_SS)	0,5230
(PMI_MANU)	0,4110
Clúster de aislamiento	
(OCUPADOS)	**0.0577**
(PIB)	**0.0556**

común ($r = 1$), el modelo absorbe la variabilidad del primer autovalor dominante ($\lambda_1^X = 3{,}68$). Como consecuencia de esto, el segundo autovalor original ($\lambda_2^X = 1{,}85$) se desplaza de forma natural hasta convertirse en el autovalor máximo de los residuos ($\lambda_{\text{máx}} = 1{,}7691$). Esta transición estructural, propia de los principios de entrelazamiento de autovalores, confirma que el algoritmo de espacio de estados ha extraído con precisión la señal macroeconómica común, dejando en la matriz de residuos únicamente el ruido sectorial remanente.

Referencias

Aarons, G. C., Caratelli, D., Cocci, M., Giannone, D., Sbordone, A. M., y Tambalotti, A. (2016, April). *Just released: Introducing the new york fed staff nowcast.* Liberty Street Economics, Federal Reserve Bank of New York. Descargado de https://libertystreeteconomics.newyorkfed.org/2016/04/just-released-introducing-the-new-york-fed-staff-nowcast/

Ahn, S. C., y Horenstein, A. R. (2013). Eigenvalue ratio test for the number of factors. *Econometrica*, *81*(3), 1203–1227.

AIReF. (2024). *Revision del modelo MIPred tras el shock del COVID-19* (Documento de Trabajo). Autoridad Independiente de Responsabilidad Fiscal (AIReF).

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*(6), 716–723.

Amengual, D., y Watson, M. W. (2007). Consistent estimation of the number of dynamic factors in a large n and t panel. *Journal of Business & Economic Statistics*, *25*(1), 91–96.

Aneiros, G. (2025). *Apuntes de la asignatura: Series de tiempo.* Material docente no publicado. (Máster en Técnicas Estadísticas (USC/UDC/UVigo))

Angelini, E., Camba-Méndez, G., Giannone, D., Reichlin, L., y Rünstler, G. (2008). *Short-term forecasts of euro area GDP growth* (Working Paper Series n.º 949). European Central Bank. Descargado de https://www.ecb.europa.eu/pub/pdf/scpwps/ecbwp949.pdf

Arencibia Pareja, A., Gómez Loscos, A., De Luis López, P., y Pérez-Quirós, G. (2024). *Modelo para la previsión del PIB de la economía española a corto plazo en tiempo real (Spain-STING): nueva especificación y reevaluación de su capacidad predictiva* (Documento Ocasional n.º 2406). Madrid: Banco de España.

Asociación Española de Economía. (2015). *Comité de fechado del ciclo económico español (CEDEC).* Descargado de http://asesec.org/CFCweb/ (Consultado el 17 de julio de 2026)

Attali, D. (2021). shinyjs: Easily perform common javascript operations in 'shiny' applications [Manual de software informático]. Descargado de https://CRAN.R-project.org/package=shinyjs (R package version 2.1.0)

Baffigi, A., Golinelli, R., y Parigi, G. (2004). Bridge models to forecast the euro area GDP. *International Journal of Forecasting*, *20*(3), 447–460.

Bai, J. (2003). Inferential theory for factor models of large dimensions. *Econometrica*, *71*(1), 135–171.

Bai, J., y Ng, S. (2002). Determining the number of factors in approximate factor models. *Econometrica*, *70*(1), 191–221.

Bai, J., y Ng, S. (2007). Determining the number of primitive shocks in factor models. *Journal of Business & Economic Statistics*, *25*(1), 52–60.

Bai, J., y Ng, S. (2008). Forecasting economic time series using targeted predictors. *Journal of Econometrics*, *146*(2), 304–317.

Bańbura, M., y Modugno, M. (2014). Maximum likelihood estimation of factor models on datasets with arbitrary pattern of missing data. *Journal of Applied Econometrics*, *29*(1), 133–160.

Bellman, R. (1961). *Adaptive control processes: A guided tour.* Princeton, NJ: Princeton University Press.

Bernanke, B. S., y Boivin, J. (2003). Monetary policy in a data-rich environment. *Journal of Monetary Economics*, *50*(3), 525–546.

Bernanke, B. S., Boivin, J., y Eliasz, P. (2005). Measuring the effects of monetary policy: A factor-augmented vector autoregressive (favar) approach. *The Quarterly Journal of Economics*, *120*(1), 387–422.

Blanco Bugueiro, A. (2024). *Modelos de inflacion: prediccion y analisis de impactos* (Trabajo Fin de Master). Master en Tecnicas Estadisticas, Universidade de Santiago de Compostela.

Box, G. E. P. (1976). Science and statistics. *Journal of the American Statistical Association*, *71*(356), 791–799.

Box, G. E. P., y Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society. Series B (Methodological)*, *26*(2), 211–252.

Box, G. E. P., Jenkins, G. M., Reinsel, G. C., y Ljung, G. M. (2015). *Time series analysis: Forecasting and control* (5.ª ed.). John Wiley & Sons.

Camacho, M., y Pérez-Quirós, G. (2011). *Spain-STING: Spain short-term indicator of growth* (Working Paper n.º 1112). Banco de España.

Chamberlain, G., y Rothschild, M. (1983). Arbitrage, factor structure, and mean-variance analysis on large asset markets. *Econometrica*, *51*(5), 1281–1304.

Chang, W., y Ribeiro, B. (2021). shinydashboard: Create dashboards with 'shiny' [Manual de software informático]. Descargado de `https://CRAN.R-project.org/package=shinydashboard` (R package version 0.7.2)

Cheng, J. (2023). promises: Abstractions for asynchronous programming [Manual de software informático]. Descargado de `https://CRAN.R-project.org/package=promises` (R package version 1.2.1)

Clements, M. P., y Hendry, D. F. (2011). *The oxford handbook of economic forecasting*. Oxford: Oxford University Press. doi: 10.1093/oxfordhb/9780195398649.001.0001

Cowpertwait, P. S. P., y Metcalfe, A. V. (2009). *Introductory time series with r*. New York: Springer Science & Business Media. doi: 10.1007/978-0-387-88698-5

Cryer, J. D., y Chan, K.-S. (2008). *Time series analysis: With applications in r* (2.ª ed.). Springer Science & Business Media.

Cuevas, A., y Quilis, E. M. (2012). A factor analysis for the spanish economy. *SERIEs: Journal of the Spanish Economic Association*, *3*(3), 311–338.

Cuevas, A., y Quilis, E. M. (2015). *Modelo integrado de predicción de la economia española (modelo MIPred)* (Documento de Trabajo n.º 2015/01). Autoridad Independiente de Responsabilidad Fiscal (AIReF).

Doz, C., Giannone, D., y Reichlin, L. (2011). A two-step estimator for large approximate dynamic factor models based on kalman filtering. *Journal of Econometrics*, *164*(1), 188–205.

Doz, C., Giannone, D., y Reichlin, L. (2012). A quasi-maximum likelihood approach for large, approximate dynamic factor models. *Review of Economics and Statistics*, *94*(4), 1014–1024.

Durbin, J., y Koopman, S. J. (2012). *Time series analysis by state space methods* (Second ed.) (n.º 38). Oxford: Oxford University Press.

Engle, R. F., y Watson, M. W. (1981). A one-factor multivariate time series model of metropolitan wage rates. *Journal of the American Statistical Association*, *76*(376), 774–781.

Eurostat. (2015). *Ess guidelines on seasonal adjustment*. Descargado de `https://ec.europa.eu/eurostat/documents/3859598/6830795/KS-GQ-15-001-EN-N.pdf`

facing, D. V., Allaire, J., Owen, J., Gromer, D., y Thieurmel, B. (2018). dygraphs: Interface to 'dygraphs' interactive time series charting library [Manual de software informático]. Descargado de `https://CRAN.R-project.org/package=dygraphs` (R package version 1.1.1.6)

Findley, D. F., Monsell, B. C., Bell, W. R., Otto, M. C., y Chen, B.-C. (1998). New capabilities and methods of the x-12-arima seasonal-adjustment program. *Journal of Business & Economic Statistics*, *16*(2), 127–152.

Friedman, M. (1937). The use of ranks to avoid the assumption of normality in analysis of variance. *Journal of the American Statistical Association*, *32*(199), 675–701. doi: 10.1080/01621459.1937.10503522

Geweke, J. (1977). The dynamic factor analysis of economic time series. En D. J. Aigner y A. S. Gold-

berger (Eds.), *Latent variables in socio-economic models* (pp. 365–383). Amsterdam: North-Holland.

Giannone, D., Reichlin, L., y Small, D. (2008). Nowcasting: The real-time informational content of macroeconomic data. *Journal of Monetary Economics*, *55*(4), 665–676.

Gómez, V., y Maravall, A. (1996). *Programs tramo and seats: Instructions for the user* (Documento de Trabajo n.º 9628). Banco de España.

Granger, C. W. J. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, *37*(3), 424–438. doi: 10.2307/1912791

Hamilton, J. D. (1994). *Time series analysis*. Princeton University Press.

Harvey, A. C. (1981). *The econometric analysis of time series*. Oxford: Philip Allan.

Harvey, A. C. (1989). *Forecasting, structural time series models and the kalman filter*. Cambridge: Cambridge University Press.

Instituto Nacional de Estadística. (2019). Estándar del ine para la corrección de efectos estacionales y efectos de calendario en las series coyunturales [Manual de software informático].

Jarque, C. M., y Bera, A. K. (1987). A test for normality of observations and regression residuals. *International Statistical Review / Revue Internationale de Statistique*, *55*(2), 163–172.

Kalman, R. E. (1960). A new approach to linear filtering and prediction problems. *Transactions of the ASME-Journal of Basic Engineering*, *82*(Series D), 35–45.

Kim, C.-J., y Nelson, C. R. (1999). *State-space models with regime switching: Classical and gibbs-sampling approaches with applications*. Cambridge, Massachusetts: MIT Press.

Krantz, S., y Bagdziunas, R. (2023). dfms: Dynamic factor models [Manual de software informático]. Descargado de `https://cran.r-project.org/package=dfms` (R package)

Kristensen, K., Nielsen, A., Berg, C. W., Skaug, H., y Bell, B. M. (2016). TMB: Automatic differentiation and Laplace approximation. *Journal of Statistical Software*, *70*(5), 1–21. Descargado de `https://www.jstatsoft.org/index.php/jss/article/view/v070i05` doi: 10.18637/jss.v070.i05

Kruskal, W. H., y Wallis, W. A. (1952). Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association*, *47*(260), 583–621. doi: 10.1080/01621459.1952.10483441

Kwiatkowski, D., Phillips, P. C. B., Schmidt, P., y Shin, Y. (1992). Testing the null hypothesis of stationarity against the alternative of a unit root: How sure are we that economic time series have a unit root? *Journal of Econometrics*, *54*(1–3), 159–178.

Lahti, L., Huovari, J., Kainu, M., y Biecek, P. (2017). Retrieving Open Government Data via R: The eurostat package. *The R Journal*, *9*(1), 385–392. doi: 10.32614/RJ-2017-019

Ljung, G. M., y Box, G. E. P. (1978). On a measure of lack of fit in time series models. *Biometrika*, *65*(2), 297–303.

Lütkepohl, H. (2005). *New introduction to multiple time series analysis*. Berlin: Springer-Verlag.

Lytras, D. P., Feldpausch, R., y Bell, W. R. (2007). Determining seasonality: A comparison of X-12-ARIMA F-tests and Q-test for seasonality. *U.S. Census Bureau Statistical Research Report*(RRS2007-01).

Mariano, R. S., y Murasawa, Y. (2003). A new coincident index of business cycles based on monthly and quarterly series. *Journal of Applied Econometrics*, *18*(4), 427–443.

Markov, A. A. (1906). Rasprostraneniye zakona bol'shikh chisel na velichiny, zavisyashchiye drug ot druga [extensión de la ley de los grandes números a cantidades que dependen unas de otras]. *Izvestiya Fiziko-Matematicheskogo Obshchestva pri Kazanskom Universitete*, *15*(4), 135–156. ((Serie 2, en ruso))

Maroz, D., Stock, J. H., y Watson, M. W. (2021, December). *Comovement of economic activity during the covid recession.* Descargado de `https://www.princeton.edu/~mwatson/papers/Covid_Factor_20211215.pdf` (Mimeo, Department of Economics, Harvard University and Princeton University)

Mincer, J., y Zarnowitz, V. (1969). The evaluation of economic forecasts. En J. Mincer (Ed.), *Economic forecasts and expectations: Analysis of forecasting behavior and performance* (pp. 3–46). National Bureau of Economic Research. Descargado de `http://www.nber.org/chapters/c1214`

Murphy, K. M., y Topel, R. H. (1985). Estimation and inference in two-step econometric models. *Journal of Business & Economic Statistics*, *3*(4), 370–379.

Newey, W. K., y West, K. D. (1987). A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica*, *55*(3), 703–708. Descargado de http://www.jstor.org/stable/1913610

Ollama Team. (2024). *Ollama: Run llama 3 and other large language models locally.* Descargado de https://ollama.com (Versión de software para ejecución local de LLMs)

Ollech, D. (2021). seastests: Seasonality tests [Manual de software informático]. Descargado de https://CRAN.R-project.org/package=seastests (R package version 0.15.4)

Ooms, J. (2014). The jsonlite package: A practical and consistent mapping between json data and r objects. *arXiv preprint arXiv:1403.2805*. Descargado de https://arxiv.org/abs/1403.2805

Pagan, A. (1984). Econometric issues in the analysis of regressions with generated regressors. *International Economic Review*, *25*(1), 221–247. Descargado de http://www.jstor.org/stable/2526118

Perrier, V., Meyer, F., y Granjon, D. (2023). shinywidgets: Custom inputs widgets for shiny [Manual de software informático]. Descargado de https://CRAN.R-project.org/package=shinyWidgets (R package version 0.8.0)

Persons, W. M. (1919). Indices of business conditions. *The Review of Economics and Statistics*, *1*(1), 5–107. doi: 10.2307/1926501

Peña, D. (2010). *Análisis de series temporales.* Alianza Editorial.

Pfaff, B. (2008). Var, svar and vec models with the vars package. *Journal of Statistical Software*, *27*(4), 1–9. Descargado de https://www.jstatsoft.org/v27/i04/

Said, S. E., y Dickey, D. A. (1984). Testing for unit roots in autoregressive-moving average models of unknown order. *Biometrika*, *71*(3), 599–607.

Sargent, T. J., y Sims, C. A. (1977). Business cycle modeling without pretending to have too much A-Priori economic theory. En C. A. Sims (Ed.), *New methods in business cycle research: Proceedings from a conference* (pp. 45–109). Minneapolis, MN: Federal Reserve Bank of Minneapolis.

Sax, C., y Eddelbuettel, D. (2018). Seasonal adjustment by x-13arima-seats in r. *Journal of Statistical Software*, *87*(11), 1–17. doi: 10.18637/jss.v087.i11

Schloerke, B., y Cheng, J. (2024). shinychat: Chat user interface for shiny [Manual de software informático]. Descargado de https://github.com/rstudio/shinychat (R package version 0.1.0)

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*(2), 461–464.

Shapiro, S. S., y Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, *52*(3–4), 591–611.

Sievert, C. (2020). *Interactive data visualization with r, plotly, and shiny.* Chapman and Hall/CRC. Descargado de https://plotly-r.com

Sims, C. A. (1980). Macroeconomics and reality. *Econometrica*, *48*(1), 1–48.

Sosa Escudero, W. (1997). Testing for unit-roots and trend-breaks in Argentine real GDP. *Economica*, *43*(1-2), 123–141.

Stock, J. H., y Watson, M. W. (1989). New indexes of coincident and leading economic indicators. *NBER Macroeconomics Annual*, *4*, 351–394.

Stock, J. H., y Watson, M. W. (2002a). Forecasting using principal components from a large multiway data set. *Journal of the American Statistical Association*, *97*(460), 1167–1179. doi: 10.1198/016214502388618960

Stock, J. H., y Watson, M. W. (2002b). Forecasting using principal components from a large number of predictors. *Journal of the American Statistical Association*, *97*(460), 1167–1179.

Stock, J. H., y Watson, M. W. (2005). *Implications of dynamic factor models for VAR analysis* (Working Paper n.º 11467). National Bureau of Economic Research. doi: 10.3386/w11467

Stock, J. H., y Watson, M. W. (2009). Forecasting in dynamic factor models subject to structural instability. En J. L. Castle y N. Shephard (Eds.), *The methodology and practice of econometrics: A festschrift in honour of david f. hendry* (pp. 173–205). Oxford: Oxford University Press.

Stock, J. H., y Watson, M. W. (2011). Dynamic factor models. En M. P. Clements y D. F. Hendry (Eds.), *The oxford handbook of economic forecasting.* Oxford University Press.

Stock, J. H., y Watson, M. W. (2016). Dynamic factor models, factor-augmented vector autoregressions, and structural vector autoregressions in macroeconomics. En *Handbook of macroeconomics* (Vol. 2, pp. 415–525). Elsevier.

Sugiura, N. (1978). Further analysts of the data by akaike's information criterion and the finite corrections. *Communications in Statistics-Theory and Methods*, *7*(1), 13–26.

U.S. Census Bureau. (2017). X-13arima-seats reference manual [Manual de software informático]. Washington, DC. Descargado de https://www.census.gov/srd/www/x13as/

Watson, M. W., y Engle, R. F. (1983). Alternative algorithms for the estimation of dynamic factor, mimic and varying coefficient regression models. *Journal of Econometrics*, *23*(3), 385–400. Descargado de https://doi.org/10.1016/0304-4076(83)90065-1 doi: 10.1016/0304-4076(83)90065-1

Welch, B. L. (1951). On the comparison of several mean values: An alternative approach. *Biometrika*, *38*(3/4), 330–336.

Wickham, H. (2024). ellmer: Call large language models from r [Manual de software informático]. Descargado de https://github.com/tidyverse/ellmer (R package version 0.1.0)

Wiener, N. (1949). *The extrapolation, interpolation and smoothing of stationary time series.* New York, NY: John Wiley & Sons.

Wijffels, J. (2023). taskscheduler: Schedule r scripts and processes with the windows task scheduler [Manual de software informático]. Descargado de https://CRAN.R-project.org/package=taskscheduleR (R package)

Wold, H. (1938). *A study in the analysis of stationary time series.* Uppsala: Almqvist & Wiksell.

Xie, Y., Cheng, J., y Tan, X. (2023). Dt: A wrapper of the javascript library 'datatables' [Manual de software informático]. Descargado de https://CRAN.R-project.org/package=DT (R package version 0.31)