Trabajo Fin de Máster

Modelo Factorial Dinámico para la Economía Gallega

Jorge C. Rella

Máster en Técnicas Estadísticas

Curso 2019-2020

II

Propuesta de Trabajo Fin de Máster

Título en galego: Modelo Factorial Dinámico para a economía galega

Título en español: Modelo Factorial Dinámico para la Economía Gallega

English title: Dynamic Factor Model for the Galician Economy

Modalidad: Modalidad B

Autor: Jorge C. Rella, Universidad de Santiago de Compostela

Directora: Rosa M. Crujeiras Casais, Universidad de Santiago de Compostela

Tutora: Belén María Fernández de Castro, ABANCA

Breve resumen del trabajo:

Desde el área de Planificación Estratégica y la Oficina de Gestión de Proyectos de ABANCA están interesados en desarrollar modelos estadísticos que permitan pronosticar la evolución de distintas variables de negocio del sistema financiero, a partir de variables macroeconómicas y de entorno. En este contexto, surge la necesidad de un modelo para obtener las proyecciones de uno de los indicadores financieros más utilizados y significativos, el Producto Interior Bruto, el cual se publica con un desfase de 50 días. Esta circunstancia sumada a la gran cantidad de variables disponibles dificultan la implementación e interpretación de sus predicciones. En este trabajo se introduce el Modelo Factorial Dinámico, con el que se intentará obtener una predicción de dicho indicador a partir de datos del entorno gallego.

Doña Rosa M. Crujeiras Casais, profesora titular del departamento de Estadística, Análisis Matemático y Optimización de la Universidad de Santiago de Compostela y doña Belén María Fernández de Castro, Coordinadora de ABANCA, informan que el Trabajo Fin de Máster titulado

Modelo Factorial Dinámico para la Economía Gallega

fue realizado bajo su dirección por don Jorge C. Rella 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 14 de julio de 2020.

La directora: La tutora:

Doña Rosa M. Crujeiras Casais Doña Belén María Fernández de Castro

El autor:

Don Jorge C. Rella

Índice general

Resumen

Resumen en español

La información de la que se dispone para el seguimiento del entorno macroeconómico suele incluir un gran número de series con distinta frecuencia temporal, con falta de información histórica y/o publicadas con retrasos considerables. Como consecuencia, la modelización debe realizarse en contextos donde los conjuntos de datos presentan bordes irregulares, datos ausentes y problemas de dimensión, dado que la sección temporal puede ser más pequeña que el número de variables.

Los modelos clásicos consideran un número de series pequeño con un histórico amplio, por lo que sufrirán problemas computacionales y de estimación en este contexto. Además, no permiten la estimación con datos ausentes, por lo que de cara a realizar predicciones estaríamos perdiendo la información más reciente. En concreto, el Producto Interior Bruto (PIB) tiene un retraso de publicación particularmente dilatado, por lo que conocer una estimación anticipada, precisa y que considere hasta el último dato de cada serie será crucial para obtener una imagen rigurosa y detallada del escenario macroeconómico de cara a la toma de decisiones. En este contexto, surgen los modelos factoriales dinámicos para solventar estos problemas, descomponiendo las dinámicas de un elevado número de series en pocos factores (o incluso solo uno) que reproducen la evolución global, considerando toda la información disponible. Los factores son inobservables y se entienden como el "estado de la economía" y reflejo de los movimientos principales de las series. Construidos de forma que condensan las dinámicas y shocks comunes, la construcción de estos factores permite que con pocas variables obtengamos una representación afín de cada serie como combinación lineal de un número reducido de componentes comunes a todas las series. De este modo, se reduce el problema de alta dimensionalidad, pudiendo emplear los factores en estimaciones y predicciones posteriores más interpretables y abordables computacionalmente.

Desde la entidad se efectúa un estrecho seguimiento de las principales variables de entorno con foco en España y en Galicia, por ser esta última el área donde ABANCA es la empresa líder en el sector bancario. El PIB ofrece una visión general de la situación económica de una región, condensando la actividad y estado económico, lo que la convierte en una serie de especial interés y motiva la necesidad de una predicción precisa. En este trabajo se empleará el Modelo Factorial Dinámico con este fin, buscando un modelo intuitivo, preciso y que se actualice conforme se dispone de más información.

English abstract

The information available for monitoring the macroeconomic environment usually includes a large number of series with different temporal frequencies, with a lack of historical information and/or published with considerable delays. Due to this, statistical modeling must be performed in contexts where data sets have irregular borders, missing data, and dimension problems, since the temporal section may be smaller than the number of variables.

Classic models consider a small number of series with a wide history, so they will suffer computational and estimation problems in this context. In addition, they do not allow estimation with missing data, so in order to make predictions we would be losing the most recent information. Specifically, the Gross Domestic Product (GDP) has a particularly long publication delay, so knowing one specified in advance, accurate and that considers every last data in each series will be crucial to obtain a rigorous and specified picture of the macroeconomic scenario for decision-making. The objective is to find a model that intuitively and precisely explains the movements of the series of interest, both to make predictions and to understand its relationship with other covariates that enter the model.

In this context, the dynamic factor model appears to solve these problems, decomposing the dynamics of a large number of series into few factors (or even just one) that reproduce global evolution, considering all the available information. The factors are unobservable and are understood as the "state of the economy" and a reflection of the main movements of the series. Constructed in a way that condenses the common dynamics and shocks, the construction of these factors allows us to obtain, with few variables, a representation of each series as a linear combination of a reduced number of components common to all the series. In this way, the problem of high dimensionality is reduced, being able to use the factors in estimates and subsequent predictions that are more interpretable and computationally approachable.

The entity closely monitors the main environment variables focused on Spain and Galicia, as the latter is the area where ABANCA is the leading company in the banking sector. GDP offers an overview of the economic situation in a region, condensing economic activity and status, making it a series of special interest and motivating the need for accurate forecasting. In this work, the Dynamic Factor Model will be used for this purpose, looking for an intuitive, precise model that is updated as more information becomes available.

Capítulo 1

Introducción

Son muchas las situaciones donde es necesario tener en cuenta varias variables simultáneamente. Por ejemplo, para estudiar el entorno macroeconómico es necesario tener en cuenta el mercado laboral, índices de negocio, de producción, de precios... En algunas ocasiones puede resultar adecuado estudiar cada una de las variables de interés de forma individual, pero en general las variables están relacionadas de tal manera que los análisis individuales suponen una pérdida de información sobre la estructura del conjunto de datos. En este contexto, sobretodo conforme aumenta el número de variables a considerar, será de interés representar la información mediante un número menor de variables, construidas como combinaciones lineales de las originales y que expliquen la mayor parte de la variabilidad original. Dentro de este tipo de modelos destacan por su uso generalizado el análisis de Componentes Principales (CP) y el Análisis Factorial.

1.1. Modelo Factorial Dinámico

El Modelo Factorial Dinámico (MFD) generaliza el Análisis Factorial, apoyándose en su filosofía, técnicas de estimación para los parámetros e interpretación. En este sentido el Análisis Factorial puede verse como un modelo de regresión que relaciona variables observadas con variables latentes (inobservables). Como se precisará más adelante, dependiendo del MFD que consideremos entrará en juego también el análisis de CP para la estimación de las variables latentes.

Se considera un conjunto de N variables observadas incorreladas $X = (X_1, \cdots, X_N)'$ y se asume que están relacionadas con r variables latentes $f_1, \cdots, f_r$ denominadas factores, donde $r < N$ mediante una relación del tipo:

$$X_i = \lambda_{i1} f_1 + \cdots + \lambda_{ir} f_r + u_i, \quad i = 1, \cdots, N$$

Los elementos λ_{ij}, denotados cargas, muestran como cada X_i depende de cada uno de los factores comunes. Cada variable u_i es particular a cada X_i, por lo que se denomina variable idiosincrática. Una de las suposiciones, y también motivación, del Análisis Factorial es que en los factores se recoge la variación principal de las series, de ahí que se suponga $E(u_i) = 0$, $\forall i$. El término $\lambda_{i1}f_1 + \cdots + \lambda_{ir}f_r$, con $i \in \{1, \cdots, N\}$ se denomina componente común, al estar conformada por factores comunes a todas las series. Es así una técnica de reducción de la dimensionalidad, buscando el mínimo de dimensiones que expliquen el máximo de información. Se diferencia del análisis de CP en cuanto a que las CP hacen hincapié en la varianza de las nuevas variables, mientras que en el Análisis Factorial interesa más explicar la estructura de las covarianzas entre variables. En el Análisis Factorial se estiman los factores a partir de un análisis por CP de la matriz de covarianza de X, por máxima verosimilitud o por un método de mínimos cuadrados. Más adelante veremos como el MFD no solo generaliza el Análisis Factorial, si no que también sus técnicas de estimación para los factores y las cargas.

Además, el MFD puede simplificar el problema de predicción para una variable de interés a partir de un vector de covariables de alta dimensión modelizando las dinámicas de cada serie en términos de un número reducido de factores latentes inobservables más un término de error de media cero correspondiente al movimiento idiosincrático de cada serie. En estos factores se supone que se recoge la información común relevante a todas las covariables. De esta forma, se simplifica el problema de regresión sin perder información sensible para las dinámicas de las series, salvo los posibles movimientos idiosincráticos propios de cada una de las covariables. Estos errores se deben a aproximar la dinámica de la variable con un número reducido de factores y a posibles movimientos que estos no puedan captar por ser muy particulares respecto a la dinámica general del resto de variables.

La idea básica es describir la dinámica estocástica de un grupo de series a través de la suma de dos componentes. El primero conformado por un grupo de factores latentes, comunes al conjunto de series y que expresan fuentes de variabilidad independientes entre sí. El segundo es idiosincrático, i.e. la parte de cada serie particular solo a ella y que por tanto no se explica por la componente común.

Una de las aplicaciones cuando se ajusta un MFD es predecir alguna variable de interés. Una vez determinados lo factores, se ajusta la relación entre la variable a predecir y los factores a través de una regresión lineal. Sea Y_t la variable sobre la que interesa realizar una predicción y, para cada $t = 1, \ldots, T$, $X_t = (X_{1t}, X_{2t}, \ldots, X_{Nt})'$ el vector de series N-dimensional de covariables, donde X_{it} representa la i-ésima variable observada en el instante t. El MFD asume que X_t admite una representación con r factores latentes, $\{f_t\}_{t=1}^T$,

$$X_t = \lambda(L)f_t + e_t, \quad t = 1, \cdots, T \tag{1.1}$$

donde f_t es un vector r-dimensional al suponer r factores, e_t es un vector N-dimensional de perturbaciones idiosincráticas de media cero. A lo largo del trabajo utilizaremos el operador retardo, L, que funciona tal que $L^k X_t = X_{t-k}$, así $\lambda(L)$ es una matriz de polinomios de retardo $N \times r$.

Una vez estimada la relación entre las covariables y los factores latentes, estos se pueden emplear para la predicción de una variable de interés, empleando toda la información de $\{X_t\}_{t=1}^T$ ahora

condensada en $\{f_t\}_{t=1}^T$,

$$Y_{t+h} = \alpha' f_t + \delta' w_t + \varepsilon_{t+h}, \quad t = 1, \cdots, T \tag{1.2}$$

donde h el horizonte de predicción, w_t un vector $m \times 1$ de variables observables (por ejemplo lags de Y_t o alguna variable que se crea que puede tener un elevado poder predictor sobre Y_t conjuntamente con f_t) y ε_{t+h} es el error de predicción. Se dispone de datos para $\{Y_t, X_t, w_t\}_{t=1}^T$ y los factores deben estimarse, al ser inobservables, a partir de $\{X_t\}_{t=1}^T$.

Con esta representación el MFD satisface los dos objetivos principales en el estudio de series económicas. El primero, la interpretación del modelo y de los parámetros, al relacionar linealmente cada serie con un pequeño número de factores comunes a todas las variables. De esta forma, es mucho más fácil encontrar patrones y relaciones entre distintas variables y así comprender las dinámicas de cada serie. La otra gran ventaja es que facilita la tarea de predicción. Por medio de una regresión simple sobre un número reducido de factores (de nuevo fácil de interpretar) se obtienen predicciones teniendo en cuenta las dinámicas e información principales de todas las series. Como indican Sargent y Sims (1977), los primeros autores que desarrollan este modelo, en vez de reducir la dimensionalidad del problema restringiendo el número de ecuaciones como en las técnicas previas, se imponen ciertas condiciones que simplifiquen el problema.

Con el MFD se solventa simultáneamente el problema de trabajar con un panel desbalanceado y el de obtener predicciones desde un enfoque multivariante para no solo un indicador de interés, si no para cualquiera que se introduzca en el modelo. Se tiene una clara reducción en la dimensión del problema, al establecer predicciones para la variable Y_t a partir de $r < N$ factores. Esto facilita también la interpretación de la serie, al tener una representación más sencilla que explique su dinámica. Cabe destacar también, que debida la relación entre los factores estimados y las N series dada por la matriz $\lambda(L)$, es posible llegar a una interpretación de qué series contribuyen más a los movimientos de Y_t. Como se ha comentado, en econometría interesa siempre la interpretabilidad de los modelos, y esta es una clara virtud del MFD.

Notemos que Y_t podría estar contenida en el vector de covariables X_t. En este caso en el vector w_t de la ecuación (1.2) no entrarían retardos de Y_t en la ecuación de predicción, ya que toda su información se supone que está capturada en los factores. De hecho, w_t suele omitirse ya que al considerar un gran número de series en X_t para la estimación de los factores, no se espera que haya información relevante más allá de estos. En algún caso puede ser de utilidad, donde se conozca de antemano una fuerte relación entre las variables. Más adelante profundizaremos más en la ecuación de predicción y en técnicas alternativas.

Cuando N es pequeño, los factores y cargas del modelo se pueden estimar parametricamente e incluir a Y_t en X_t, pero al considerar un mayor número de series se pierde justificación para este enfoque debido al aumento en el número de parámetros. El MFD puede ajustarse desde un enfoque tanto paramétrico como no paramétrico. En los primeros trabajos, se solía inclinar por el segundo, al contar con un número reducido de variables y no necesitar la selección de ningún parámetro previo.

Los trabajos más recientes sostienen que suponer algún tipo de estructura sobre los parámetros del modelo conduce a estimaciones más eficientes y precisas. Debido a las ventajas e inconvenientes de cada uno, también han aparecido técnicas mixtas, donde se conjugan ambos enfoques. A lo largo del trabajo se introducen diversas técnicas de estimación y las ventajas y desventajas de cada enfoque.

Sean $\{y_t\}_{t=1}^T$ y $\{x_t\}_{t=1}^T$ realizaciones en $t = 1, \ldots, T$ de X_t e Y_t respectivamente. Podría pensarse en considerar simplemente la regresión de $\{y_t\}_{t=1}^T$ sobre $\{x_t\}_{t=1}^T$ para realizar las predicciones como en el enfoque tradicional, pero resulta inadecuado ya que el término de error es proporcional a N/T, por lo que en el marco donde nos encontramos no tenderá a reducirse. En un escenario como el económico donde las series tienden a no ser suaves, será difícil o imposible muchas veces llegar a una solución satisfactoria. En caso de contar con un número reducido de series este error se puede considerar despreciable, pero conforme aumenta N el error cuadrático medio (MSE, del inglés Mean Square Error) del ajuste no se reduce. Al estar en un escenario donde interesa trabajar con un numero de series grande, la motivación para el MFD está clara. Es análogo al clásico balance entre sesgo y varianza de un modelo. Otra alternativa podría ser utilizar algún criterio de información para seleccionar un subconjunto de las N variables, como el de Akaike (AIC, del inglés Akaike Information Criterion) o el de Bayes (BIC, del inglés Bayes Information Criterion). El problema de estos criterios es que ante un N elevado el número de modelos a evaluar crece drásticamente, por lo que computacionalmente no serán abordables.

En lo que sigue, se asume que las variables observables y las latentes son estacionarias e integradas de orden cero ($I(0)$), sin raíces unitarias. En caso de considerar algún modelo autorregresivo con raíz unitaria estaríamos modelando las series como un modelo en el que cualquier shock de la variable tiene un efecto permanente en la serie. Como indica Sosa-Escudero (1997), esto iría en contra del enfoque tradicional, donde las variables económicas pueden ser caracterizadas como fluctuaciones estacionarias alrededor de una tendencia determinista, y por tanto los shocks tienen efectos pasajeros. También se supone que las series han sido estandarizadas de forma que tengan desviación típica unitaria. En la práctica, en primer lugar se diferencia cada serie tantas veces como sea necesario hasta eliminar la tendencia y a continuación se pueden eliminar posibles tendencias o movimientos de baja frecuencia con el uso de algún filtro. La presencia de raíces unitarias se estudia con algún test como el de Dickey-Fuller (véase Fuller (1996)). También se acostumbra a eliminar la posible presencia de atípicos para que no influyan en la estimación de los factores. Un criterio, por ejemplo, es eliminar las observaciones que superen 4 veces el rango intercuartílico desde la mediana y sustituirlas por valores perdidos según sugieren Giannone *et al.* (2008).

Al considerar un elevado número de series cabe la posibilidad de que algunas sean cointegradas. Esto significa que comparten una tendencia estocástica común. Formalmente, se dice que dos series $\{w_t\}_{t=1}^T$ y $\{z_t\}_{t=1}^T$ están cointegradas si existen $a, b \in \mathbb{R}$ tal que $aw_t + bz_t = e_t$ es una serie estacionaria. Esta fórmula se conoce como ecuación de cointegración. Si algunas de las variables en el conjunto de datos X_t son cointegradas, diferenciarlas podría hacer que se pierda cierta información que podría estar presente en los residuos de la ecuación de cointegración. Para tratar con series cointegradas se puede incluir la primera diferenciación de una parte de las variables y las series corregidas con el término

del error de las restantes. Esto es importante si se cree que este residuo de la regresión puede tener información para la estimación de alguno o varios de los factores. Otra alternativa es incluir las series diferenciadas sin atender al término de error que se podría perder con esta aproximación. Este enfoque es apropiado cuando no cabe esperar que influya la posible pérdida de información. Suele utilizarse cuando estamos ante variables macroeonómicas, donde no se espera que influyan estos términos. La cointegración se puede estudiar con algún test como el de Johansen (véase Johansen (1988)).

1.2. Motivación en econometría del modelo

Desde mucho antes de la aparición del MFD en la literatura, los economistas han advertido las dinámicas comunes existentes en los mercados y variables económicas, en cuanto a que tienden a "subir" y "bajar" al mismo ritmo y con la misma intensidad. Esto se magnifica cuando las variables tienen algún tipo de relación, como formar parte del mismo sector o de la misma región. El Análisis Factorial como tal requiere de ciertas modificaciones para adecuarse al contexto de las series de tiempo. Esta supone la primera motivación para la aparición del MFD, que busca implementar la filosofía intrínseca a los factores latentes al contexto de las series temporales. Los primeros trabajos donde aparece el MFD sostienen que estos movimientos síncronos son fruto de un elemento común dirigido por una variable subyacente inobservable, interpretada como el "estado de la economía", en línea con la intuición de la existencia de dinámicas comunes dentro del entorno económico. La idea primitiva fundamenta que las variables económicas tienen una dinámica común conducida por una variable latente inobservable, en línea con el Análisis Factorial. Esto explicaría las subidas y bajadas simultaneas de muchas series económicas y los momentos de volatilidad que tienden a ocurrir también en instantes similares. Este "estado de la economía" se interpreta que es lo que captura el MFD cuando se considera un solo factor. De esta forma, el MFD se alinea con la teoría y observación empírica de las variables económicas y dota al investigador de una herramienta fácil de interpretar y predecir de cara al futuro.

Otra de las motivaciones para emplear el MFD en el contexto económico es el gran número de series disponibles. La dimensión del problema de predicción e interpretación de las variables económicas ha ido creciendo de forma exponencial desde mediados del siglo pasado, cuando se comienza a comprender la importancia de los datos para entender y modelar los movimientos económicos. A lo largo de las últimas décadas se han generado una gran cantidad de datos para variables macroeconómicas y financieras, generalmente de manera mensual y trimestral. En este marco se tiene el problema de disponer de cientos o inclusos miles de series pero con un número de observaciones reducido, por ejemplo 20-40 años de datos trimestrales. Por tanto, debemos disponer de un modelo que permita interpretar y predecir las distintas variables cuando el número de series excede el número de observaciones. En econometría uno de los propósitos principales de todo estudio es la interpretación de los resultados, lo cual se dificulta debido al elevado número de variables a considerar. Esta es una de las principales motivaciones para los MFDs en este contexto, al condensar la información de las series en un número muy reducido de factores permitiendo una interpretación rápida e intuitiva.

A lo largo del trabajo se emplea la expresión "técnicas clásicas" debido a su utilización histórica y por su prevalencia en el entorno actual. Las publicaciones más destacadas en MFD se consideran contemporáneos en cuanto a que la mayor parte de su desarrollo y aplicación se ha dado en el Siglo XXI. Aunque la primera introducción de esta clase de modelos se tiene en Sargent y Sims (1977), trabajos como Stock y Watson (2002), Stock y Watson (2002a), Bai y Ng (2008), Mariano y Murasawa (2003) o Bai (2003), entre otros, son los principales desarrolladores del MFD y las principales referencias para su aplicación práctica. De hecho, es tal la novedad, que la implementación práctica en este trabajo se realiza con un paquete publicado en 2012 (véase Holmes *et al.* (2012)). Hoy en día se vive en el "boom" del Big Data y se debe ajustar la metodología a un contexto donde el número de variables supera al de observaciones. Dentro de los métodos clásicos para el modelado de series de tiempo destacan los modelos autorregresivos de media móvil (ARMA, por sus siglas en inglés). Este modelo interpreta y ajusta una serie de tiempo en base a observaciones pasadas de la propia serie más términos de ruido blanco (variables aleatorias i.i.d de media cero). La desventaja de este modelo es que no introduce información auxiliar ajena a la serie, lo cual es difícil de justificar en un contexto en que se disponen de hasta miles de variables. A raíz de esto el modelo ARMA Vectorial (VARMA, por sus siglas en inglés) es quizás el más utilizado, ya que es una generalización del modelo ARMA al marco vectorial, incluyendo en el ajuste de cada serie el efecto de las demás covariables consideradas en el modelo. El problema es que el número de series que puede modelar es muy reducido, del orden de 10, debido a que el número de parámetros hace crecer el problema computacionalmente y conlleva a dificultades para estimar los parámetros del modelo, como problemas de convergencia o matrices mal condicionadas.

Las técnicas y modelos clásicos para trabajar con series de tiempo no resultan adecuados o no son abordables computacionalmente si se considera un número elevado de series. Estos, suponen que el número de series no supera el de instantes temporales y se modelan como procesos autorregresivos, como el modelo Vectorial Autoregresivo (VAR). Los modelos construidos con pocas series están expuestos a sufrir pérdidas de información (y un consecuente sesgo en las estimaciones) y a que un shock en una serie tenga un peso excesivo, afectando en las estimaciones y predicciones posteriores. Por consiguiente, se busca un modelo capaz de explicar las dinámicas de las variables incluyendo todas las series.

Otro problema a la hora de trabajar con variables macroeconómicas es el retraso en la publicación de muchas de las variables y la consecuente pérdida de información al considerar medidas agregadas de la actividad económica. Los retrasos, particulares de cada serie, implican contar con conjuntos de datos con bordes irregulares, en el sentido de que en cada instante se conocen los datos actualizados para algunas variables y para otras no. Especialmente sensible es el retraso asociado a los agregados macroeconómicos, como el PIB, que condensa información relevante sobre el estado de la economía pero que no se conoce hasta pasados 50 días desde el final del trimestre. Dentro de las series de interés aparecen series mensuales e incluso trimestrales, pero en las instituciones financieras es necesario monitorizar la actividad económica en una base hasta diaria, por lo que interesa obtener predicciones con esa regularidad. Como en toda predicción, interesa considerar toda la información disponible hasta el momento. La regularidad dispar de las covariables (desde diarias a trimestrales o anuales) combinado con el retraso heterogéneo en la publicación de cada variable es un problema para las técnicas clásicas como el VAR, fruto de los datos ausentes dispuestos de forma irregular.

En este contexto interesa contar con un modelo que actualice sus estimaciones cada día conforme se van conociendo nuevos datos, sin renunciar a ninguno. De esta forma, se espera que a cada dato nuevo que se conozca en el momento, este entre en el ajuste del modelo y la predicción mejore respecto a la del día anterior. Los agregados macroeconómicos más importantes tienen una frecuencia muy baja, por lo que será de interés conocer y predecir el movimiento a más corto plazo considerando toda la información disponible hasta el momento. Las técnicas clásicas no permiten trabajar con bordes dentados, lo que implica que no se puede hacer una predicción "al día". La expresión bordes dentados es muy recurrente cuando se habla de los paneles de datos en el MFD, ya que hace referencia al fenómeno de considerar series con fechas de publicación asíncronas, lo que supone en un instante de tiempo tener algunas series actualizadas y otras con una o varias medidas ausentes. Debemos considerar el subconjunto de los datos más grande para el que disponemos de un borde completo, lo que conlleva una perdida de información. También debemos afrontar el problema de que no todas las series disponen de datos desde el mismo instante. Algunas variables cuentan con datos de muchos años mientras que otras cuentan con apenas $10-20$ años de historia. Esto provoca también la aparición de datos ausentes "al principio" del conjunto de datos si queremos contar con toda la información disponible. Los métodos clásicos imponen tener que sustituir los "huecos" con algún valor (media, mediana, ceros, repetir el valor del último mes, ...) y/o cortar el número de series e instantes temporales en aras de contar con un conjunto de datos sin valores ausentes. Esta es una de las principales motivaciones que ha llevado a numerosas instituciones a emplear el MFD, ya que consideran toda la información disponible hasta la fecha, introduciendo todos los datos conocidos en cada momento, sin importar que no se conozcan los datos actualizados para todas las variables o los posibles "huecos" en las series derivados de lo comentado anteriormente.

1.3. Conceptos previos

En la literatura del MFD, se suele llamar al conjunto de datos panel. Este concepto no es más que la idea de agrupar por columnas todas las series de las que disponemos (cada variable ocupa una columna y el tiempo se mueve por filas) expresadas en la mayor frecuencia dentro del panel, creando una matriz susceptible de tener valores ausentes debido a la medición a distinta frecuencia de cada serie, por empezar en distinto instante o por la fecha dispar de publicación del último dato. Cuando en el momento de la estimación no se dispongan de todas las series actualizadas se dirá que el panel tiene los bordes dentados, y cuando haya algún valor ausente se dirá que está desbalanceado.

A lo largo del trabajo se introducirán técnicas para el ajuste del modelo cuando se disponga de todas las observaciones para todas las variables (panel balanceado) y para cuando tengamos datos ausentes para alguna(s) (panel desbalanceado). En la Figura 1.1, inspirada en el panel representado en Cuevas y Quilis (2012), se muestra un ejemplo visual de un panel desbalanceado. Se puede observar que no todas las series cuentan con información desde el mismo instante y que algunas no cuentan con el último dato actualizado debido a su retraso en la publicación. El panel también es susceptible de tener datos ausentes en su interior debido a considerar variables con distinta frecuencia. En el

marco clásico, se selecciona uno de los subpaneles balanceados de la Figura 1.1 y se procedería con el ajuste del modelo considerado. Por ejemplo, en un modelo VAR se estimarían los parámetros del modelo en base al panel longitudinal o al panel balanceado desde la sección cruzada, pero no se podría considerar el panel completo. Como indican Cuevas y Quilis (2012), en el primer caso se descarta un número relevante de series, lo que reduce la precisión y el poder predictor del modelo. En el segundo, el número de observaciones es muy pequeño, haciendo el horizonte de predicción demasiado grande.

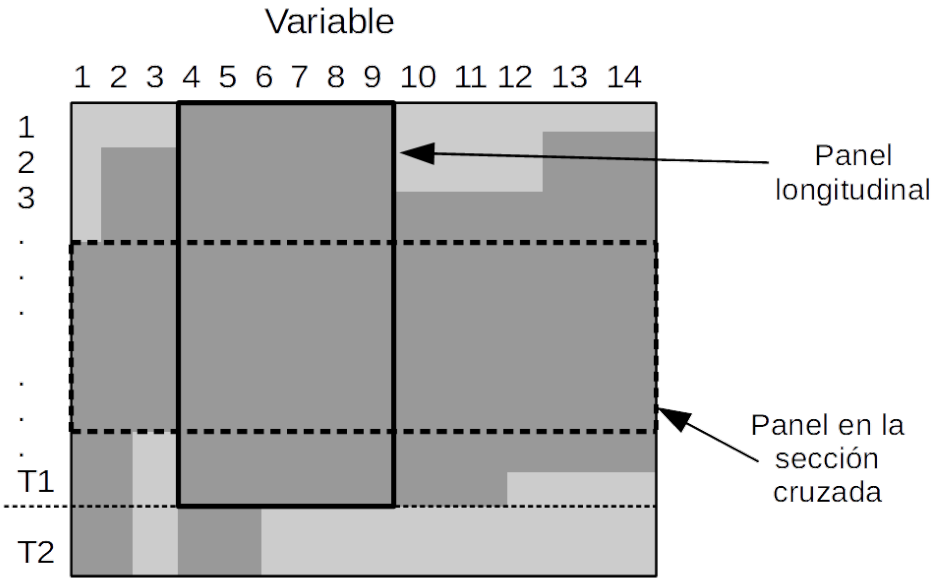

Figura 1.1: Panel de datos desbalanceado, representación minimalista inspirada en el de Cuevas y Quilis (2012).

Forecasting, nowcasting y backcasting Finalmente, se deben puntualizar los distintos tipos de predicciones que pueden ser de interés en el contexto que nos ocupa. En concreto, podemos tener que realizar *forecasting*, *nowcasting* o *backcasting*, según el momento en que se realice y para qué instante consideremos la predicción. El término *forecast* se refiere a predicciones a más largo plazo que el propio instante en el que nos encontramos, i.e. predicciones a futuro. Se denomina *nowcasting* cuando se considera la predicción respecto al instante actual, mientras que *backcasting* se refiere a la predicción de un dato del que no se dispone en períodos pasados. Emplearemos estos anglicismos a lo largo del trabajo ya que en la literatura es como se refieren generalmente. Por ejemplo, para el caso del PIB, interesa conocer una estimación mucho antes de su publicación y no un mes y medio más tarde, cuando se conoce el dato real. Supongamos que queremos obtener una estimación del PIB para el segundo trimestre de 2020 (desde abril a junio de 2020). Si el ejercicio se realiza durante el segundo trimestre, por ejemplo en junio, la predicción se clasifica como *nowcasting* y si la consideramos antes del segundo trimestre, en febrero por ejemplo, *forecast*. Finalmente, si estamos en un instante posterior al segundo

trimestre pero el dato real para el PIB no se conoce todavía, situación que se da en julio y principios de agosto en nuestro ejemplo, se clasifica como *backcasting*. En la Figura 1.2 se representan los tres tipos de predicción de forma visual para la predicción del dato del PIB para el segundo trimestre del año, teniendo en cuenta que el retraso en la publicación es de 50 días desde el final del trimestre.

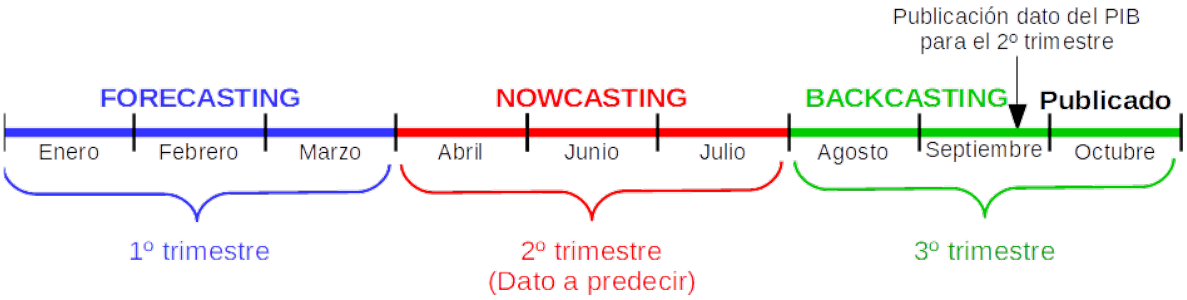

Figura 1.2: Expresiones para la predicción del PIB del 2º trimestre, publicado con un retraso de 50 días. Distinción entre los períodos a los que hacen referencia forecasting, nowcasting y backcasting.

PIB La macroeconomía se ocupa de estudiar el funcionamiento de la economía en su conjunto, siendo uno de los objetivos agregar los distintos bienes y servicios hasta reducirlos a un solo bien genérico. El más importante de los agregados es el PIB, que se define como el valor monetario total de los bienes y servicios producidos para el mercado durante un año dentro de las fronteras de un país. Debido a su clara importancia para el modelado y entendimiento del escenario económico de una región, la variable de interés principal en este trabajo será el PIB de Galicia[1]. En el ámbito nacional, es el Instituto Nacional de Estadística (INE)[2] el organismo encargado de su publicación, y en el caso de la comunidad autónoma de Galicia es el Instituto Galego de Estatística (IGE)[3]. El PIB es una de las principales macromagnitudes de síntesis, de carácter coyuntural, cuyo objetivo es proporcionar una descripción cuantitativa y coherente de la evolución reciente de la economía. De acuerdo con su definición en la contabilidad nacional, el PIB se define como el valor de todos los bienes y servicios nacionales (se excluyen aquellos bienes de carácter intermedio con el fin de evitar una doble contabilización) producidos en un territorio durante un período de tiempo determinado, generalmente un año. Cabe destacar, que el PIB no mide la riqueza o patrimonio de un país, sino su capacidad productiva, es decir, refleja la capacidad de una economía para producir riqueza a lo largo de un período.

Queda clara la importancia que adquiere entonces el PIB de cara a entender y predecir el contexto maroeconómico en el que se encuentra la entidad a fin de tomar decisiones consecuentes con ello. Cabe destacar que el PIB es una serie sintética resultado de combinar diversos indicadores de la producción

[1] http://www.ige.eu/estatico/estat.jsp?ruta=html/gl/OperacionsConxunturais/ContabilidadeTrimestral.html
[2] https://www.ine.es/
[3] https://www.ige.eu/web/index.jsp?idioma=es

y actividad económica, por lo que de esta forma, el PIB es equivalente a una componente común salvo por no estar construido de tal forma. Como resultado, se espera que el MFD pueda modelar la dinámica de esta variable y mejorar las estimaciones realizadas por los métodos clásicos.

1.4. Revisión en la literatura

El MFD, introducido en Sargent y Sims (1977), ha sido posteriormente desarrollado en trabajos destacados como Stock y Watson (1998), Stock y Watson (2002), Stock y Watson (2002a) entre otros recogidos en la bibliografía, donde se han proporcionado numerosas herramientas para la estimación y aplicación del modelo, así como modelos explotando la filosofía tras los factores y mostrando el buen comportamiento en la práctica. Bai y Ng (2002) ofrecen criterios de información para la selección del número de factores y Bai y Ng (2008) teoría sobre la distribución y comportamiento asintótico de los factores. Todo esto ha dotado a numerosos autores con las herramientas necesarias para la aplicación y desarrollo de los MFDs en diversos escenarios.

Numerosos autores han aplicado el MFD obteniendo buenos resultados para predecir agregados macroeconómicos como Camacho y Pérez-Quiros (2009) para el PIB español. Hay una gran evidencia empírica de que el MFD con un número pequeño de factores captura la mayor parte de las dinámicas de las series macroeconómicas tras el periodo de guerras en EEUU. Sargent y Sims (1977) concluyen que dos factores explican el 80 % o más de la mayoría de las variables económicas, incluyendo el desempleo, producción, inflacción,... Stock y Watson (2002a) lo aplican al índice de producción industrial mejorando el MSE de los métodos clásicos. Giannone *et al.* (2008) aplican el MFD para predecir el PIB con notables resultados. Numerosos artículos muestran también el gran poder predictor de un número reducido de factores (del orden de 2) para los movimientos de los agregados macroeconómicos más destacados. Dentro del marco español un trabajo destacado es el de Doz *et al.* (2011), donde justifican una selección de variables económicas para la aplicación del MFD y los distintos pasos que siguen para su ajuste. Hacen hincapié en la importancia de considerar un panel de datos con un verdadero poder explicativo sobre la variable de interés, obteniendo un ajuste para el PIB español con gran precisión y capaz de predecir cambios repentinos y/o bruscos en su nivel.

Otros autores emplean el MFD para concentrarse únicamente en los factores estimados, empleándolos como índices de alguna variable latente (lo cual es consecuente con la definición y estimación que se realiza de los factores). Melo *et. al* (2005) construyen un índice de percepción del riesgo, en donde en vez de intentar predecir ninguna variable, buscan construir una a modo de factor cuyos niveles indiquen el "sentimiento" que hay entre los inversores relativo al riesgo en los bonos de un país. En este trabajo no se busca predecir ninguna variable. El interés es obtener una nueva variable latente que a partir de un conjunto grande de series pueda explicar el nivel de riesgo que se percibe en el mercado de un país. Camacho y Doménech (2012) aplican el MFD para obtener una variable cuyo nivel indique (y prediga a futuro) si la economía se encuentra en recesión. Estos trabajos ejemplifican otra de las aplicaciones del MFD, donde el investigador se centra en los factores en aras de entender los

movimientos, en muchos casos interesando solo la dirección que tomará el aglomerado de las variables a estudiar. De esta forma, se toman los factores estimados como un resumen intuitivo del cúmulo de variables que facilita la interpretación y predicción del mercado o economía subyacente.

1.5. Estructura del trabajo

El objetivo de este trabajo es doble. En primer lugar se pretende realizar una revisión que presente de forma clara y con un cierto grado de profundidad la filosofía y metodología asociada al MFD, introduciendo los distintos enfoques, técnicas de estimación, propiedades y aplicaciones del modelo. Asimismo, se comentan las herramientas disponibles para implementar el MFD y las funciones de las que se dispone en el software R (R Core Team (2018)). El segundo propósito es el de proporcionar a ABANCA Corporación Bancaria S.A. un modelo detallado y ya implementado para la predicción del principal medidor de la economía, el PIB, que permita a la entidad contar con proyecciones a futuro y a presente precisas considerando un gran número de variables. En el Capítulo 2 se introduce el MFD con detalle, precisando todos los elementos que conforman el modelo, las dos formas que puede tomar dependiendo del enfoque que se haga de los factores y la correspondiente notación. Se muestra la equivalencia entre las dos formas y los motivos que llevan a considerar cada una. También se precisa la definición para el MFD considerado según se satisfagan las hipótesis de homocedasticidad e independencia de los errores. Con el modelo ya introducido, en el Capítulo 3 se repasan todas las técnicas de estimación, las cuales pueden clasificarse como paramétricas, no paramétricas o mixtas. Se mostrará los distintos enfoques disponibles según la forma del MFD considerada y los datos ausentes en el panel. Asimismo, de cara a cerrar el tratamiento del panel, se indica cómo manipularlo en caso de contar con variables medidas en frecuencias heterogéneas y se recogen varios criterios para la selección de las variables que entran en el modelo. El Capítulo 4 recoge por separado las técnicas para la estimación del número de factores latentes. La estimación de los parámetros y ajuste del modelo posteriores se basan en el número de factores estimados, por lo que este será un tema clave a fin de construir un modelo adecuado. Una vez estimados todos los parámetros relativos al modelo, en el Capítulo 5 se recogen varios resultados sobre la distribución y las componentes comunes, así como tasas de convergencia. Finalmente, se indica la construcción de intervalos de confianza en base a estos resultados y también a partir del método bootstrap. En el Capítulo 6, considerado el ajuste del modelo según los capítulos anteriores se detalla el procedimiento de predicción en base al MFD para la variable de interés y se introduce un modelo VAR en donde los factores entran como covariables. Los últimos capítulos pueden considerarse prácticos. En el Capítulo 7 se recogen los dos paquetes disponibles para el ajuste de MFDs, indicando las técnicas y funciones disponibles en cada uno, así como sus respectivas ventajas y desventajas. El Capitulo 8 recoge un estudio completo empleando el MFD para predecir la variable de interés del trabajo, el PIB. En este capítulo se implementa, siguiendo la estructura del trabajo un MFD completo, desde el tratamiento del panel de datos hasta la regresión del PIB sobre los factores extraídos del panel, pasando por la estimación del número de factores y de los propios factores. Se motiva la selección de las distintas variables consideradas para el análisis,

tomando como panel inicial el conjunto de indicadores que se monitoriza desde el área de Estudios de la Entidad. También se incluye un estudio detallado de las variables y el factor extraído, de forma que se obtenga una visión clara de qué se está haciendo y se afiance el entendimiento del MFD. Se calculan y representan también los intervalos de confianza para la predicción, que se introducirán previamente en el Capítulo 5. Finalmente, se realiza un pequeño ejercicio de simulación en donde se repite todo lo anterior considerando el panel de datos tal y como se tendría (teniendo en cuenta los retrasos de publicación) entre julio de 2019 y abril de 2020 para la predicción del dato para el PIB del 1^{er} trimestre de 2020. Con este ejercicio se pretende mostrar cómo varía la predicción que se obtiene con el MFD conforme se conoce más información y examinar los tres tipos de predicción introducidos.

Capítulo 2

Modelo Factorial Dinámico

El MFD representa la evolución de un vector N-dimensional de series temporales observables, $\{X_t\}_{t=1}^{T}$, en términos de un número reducido (r) de factores latentes inobservables comunes a todas ellas, representados como un vector r-dimensional f_t, más un término de error y/o dinámica idiosincrática de la propia serie, e_t. Los factores evolucionan a lo largo del tiempo según algún proceso autorregresivo y se supone que captan la mayor parte de la información relevante relativa al movimiento de las N series temporales observadas. Los factores sintetizan los movimientos principales comunes a todas las series y los condensan en un número reducido de variables, inobservables empíricamente. Por medio de una matriz de parámetros, denominados cargas, expresaremos cada serie como una combinación lineal de los r factores y posibles lags de estos. Debido a esta reducción de la dimensión, se tiene cierta pérdida de información relativa a movimientos particulares de cada serie, que los factores no pueden captar al no formar parte de las dinámicas comunes del resto de variables. Este error producido por la aproximación con r factores se conoce como perturbaciones idiosincráticas, y se supone de media cero para cada serie. Un gran interés en econometría, como se ha comentado, es la interpretabilidad, por lo que esta pérdida de información es justificable en aras de obtener un modelo más entendible. Además, gracias a la proyección sobre r factores, se facilitan las tareas de predicción tanto computacionalmente como en términos de entendimiento; tareas que bajo otro modelo podrían no ser ni aplicables.

En la Figura 2.1 mostramos el ajuste del MFD con $r = 1$ factor sobre un conjunto de datos con $N = 10$. Puede observarse como las principales dinámicas de todas las series se captura en el factor, coincidiendo los momentos de subidas, bajadas, mayor y menor volatilidad de las variables observadas y el factor latente estimado. Los movimientos "extremos" en algunos puntos de alguna de las series no son captados por el factor al disgregarse del resto de las series. Este tipo de dinámicas, propias de una sola variable, son las que recoge la componente idiosincrática e_t. El factor procura captar los movimientos comunes a todas las series, y los movimientos de estas componentes idiosincráticas se "pierden" en el factor en aras de capturar la dinámica principal. En la Figura 2.1 se pretende recoger

las dos virtudes expuestas hasta ahora del MFD. Notemos que reduciendo la dimensión de los datos a una sola serie, disponemos de una variable latente que captura la dinámica principal, permitiendo la interpretación más sencilla de los movimientos de las series. Con solo una variable, se entienden de forma más intuitiva los momentos de volatilidad o de crecimientos y decrecimientos. En el ejemplo solo se consideran 10 series, pero nótese que conforme crece N la interpretación conjunta de todas las series se dificulta hasta el punto de ser imposible sin una reducción de la dimensionalidad. La otra ventaja del MFD comentado es cómo facilita el forecast. Con la información más relevante de todas las series condesada en una sola variable, si interesase obtener la predicción de alguna variable considerando toda la información disponible en el panel ahora resulta mucho más fácil efectuar algún tipo de regresión o proyección hacia el futuro de la variable de interés a partir del factor estimado. Igualmente, al tener relacionadas las covariables del panel con el factor mediante las cargas, la predicción para las series del panel también se simplifica en la misma medida. Además, al disponer de la relación entre los factores y las variables por medio de las cargas, como se indica en (1.1), una vez realizado el forecast es de nuevo más fácil la interpretación de los resultados, pudiendo interpretar qué variables influyen en mayor medida en la composición del factor y en consecuencia sobre las demás variables. En la Figura 2.1 se muestra en rojo la proyección a 2 pasos del factor (más adelante se explica como se efectúa con detalle). Como primera interpretación puede deducirse que el global de las series tenderá a descender. Una vez calculado el forecast para el factor, podría establecerse el forecast para cualquiera de las variables empleando la relación dada por las cargas.

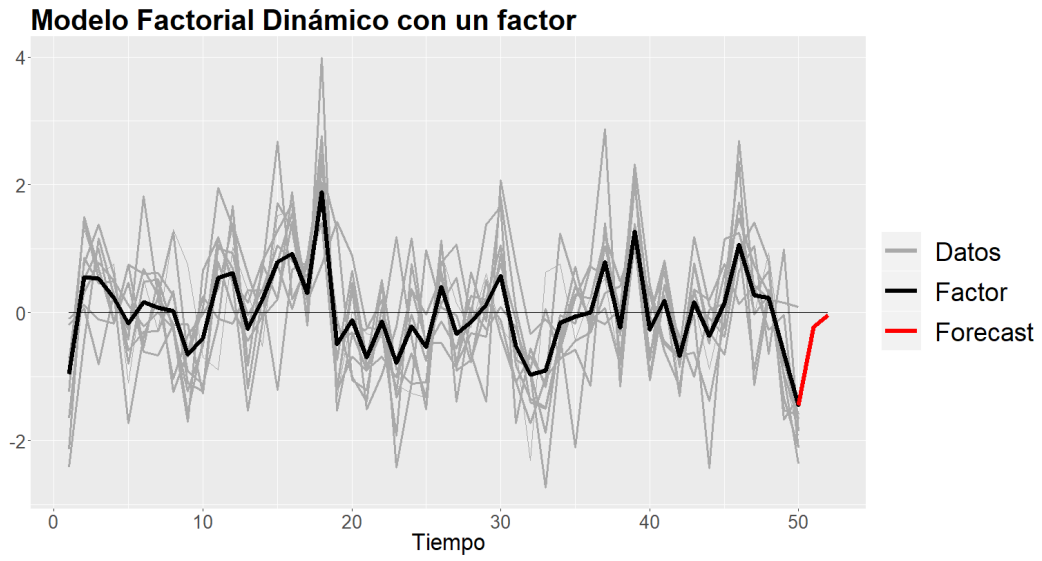

Figura 2.1: Modelo Factorial Dinámico ajustado con 1 factor sobre 10 series simuladas. Con líneas grises se representan las 10 series simuladas, en negro el factor latente estimado a partir del panel de 10 variables y en rojo el *forecast* a 1 y 2 pasos para el factor.

Existen dos formas alternativas para representar el MFD. La forma dinámica expresa la dependencia de X_t en función de los factores (y posibles retardos de estos) de forma explícita. La forma estática

representa estas dinámicas de forma implícita. Dependiendo de la representación que hagamos del modelo, llegaremos a distintas técnicas para la estimación de los factores. En los primeros trabajos, como en Stock y Watson (2002), suele considerarse el modelo estático al ser más conveniente para la estimación de los factores desde un enfoque no paramétrico y permite considerar un mayor número de series. Al aumentar N, crece el número de parámetros para proceder paramétricamente y se toma un enfoque no paramétrico, y en caso de tener numerosos datos ausentes, la representación estática es más conveniente para lidiar con ese tipo de paneles.

Se puede representar el MFD dentro de los procesos de espacio de estados, en los cuales se expresan las variables observables en términos de componentes latentes (inobservables) que evolucionan según algún tipo de dinámica de retardos finita. La representación del MFD en un modelo de este tipo motivará el enfoque paramétrico para la estimación de los factores, apoyándose en las técnicas ya desarrolladas para estos modelos.

2.1. Forma dinámica del Modelo Factorial Dinámico

La forma dinámica del MFD es la representación más intuitiva para el modelo, al reflejar explícitamente las relaciones entre las variables observables y los factores. Es la "traducción" directa del MFD a ecuaciones a partir de las variables como series de tiempo y su expresión es análoga a la de un modelo de espacio de estados, interpretando la primera ecuación la de observación y la segunda la de estados.

La premisa del modelo es que un número reducido (r) de factores dinámicos, f_t, conducen el movimiento de un vector de series de tiempo N-dimensional, X_t, al cual también afecta un vector de media cero de perturbaciones idiosincráticas relativas a cada serie, e_t. Estas perturbaciones provienen del posible error al aproximar el movimiento de una serie temporal a partir de un número reducido de factores y por comportamientos particulares de la propia serie que los factores no logran captar. Los factores siguen algún proceso estocástico, habitualmente tomado como un modelo autorregresivo (VAR). La forma dinámica del MFD es:

$$X_t = \lambda(L) f_t + e_t, \quad t = 1, \cdots, T \tag{2.1}$$

$$f_t = \Psi(L) f_{t-1} + \eta_t, \quad t = 1, \cdots, T\,, \tag{2.2}$$

donde se contemplan N series, por tanto X_t y e_t son de dimensión $N \times 1$ y hay r factores, por lo que f_t y η_t (el error del proceso VAR) son vectores r-dimensionales. L indica el operador retardo, $\lambda(L)$ y $\Psi(L)$ son polinomios de retardo (de grado a lo sumo $p < \infty$) en forma matricial de dimensión $N \times r$ y $r \times r$ respectivamente. El i-ésimo polinomio de retardo $\lambda_i(L)$ (i-ésima fila de $\lambda(L)$) se denomina carga del factor para la i-ésima serie, $\{X_{it}\}_{t=1}^{T}$, y $\lambda_i(L) f_t$ la componente común de la i-ésima serie. Se asume que los procesos (2.1) y (2.2) son estacionarios. También se asume que las componentes idiosincráticas son incorreladas con todos los factores en todos los rezagos, i.e. $E[e_t \eta'_{t-k}] = 0 \ \ \forall k$.

El error del proceso VAR de los factores se representa como:

$$\eta_t = Bu_t, \quad u_t \sim i.i.d.N(0, I_r) \ \ y \ \ B \in \mathcal{M}_{r\times r} \tag{2.3}$$

La componente idiosincrática, e_t, puede estar serialmente correlada. En caso de suponer algún tipo de dinámica de este tipo para e_t, el modelo (2.1) y (2.2) no está completamente especificado y no se puede abordar. En sucesivas técnicas para la estimación, como en el modelo de espacio de estados, se establece un modelo paramétrico para la perturbación idiosincrática asociada a la i-ésima serie, $\{e_{it}\}_{t=1}^{T}$, para especificar por completo el modelo. Por ejemplo una estructura $AR(p')$:

$$e_{i,t} \ = \ \delta_i(L)e_{i,t-1} + \epsilon_{i,t} \ = \ \sum_{j=1}^{p'} \delta_{ij} e_{i,t-j} + \epsilon_{i,t}, \ \ \epsilon_{i,t} \sim N(0, \sigma_i^2) \ \ i.i.d., \quad i = 1, \cdots, N \tag{2.4}$$

Modelo Factorial Dinámico exacto En el MFD exacto se supone que las componentes idiosincráticas son mutuamente incorreladas entre ellas y en todos los retardos, i.e. $E[e_{it}e_{js}] = 0 \ \ \forall s, t$ si $i \neq j$. Bajo este modelo la correlación de una serie con el resto se produce únicamente a través de los factores f_t. Es decir, supuestos los factores, el resto de series $\{X_{jt}\}_{t=1}^{T}$ no tiene ningún poder explicativo sobre $\{X_{it}\}_{t=1}^{T}$, para $j \neq i$. Esto se debe a que bajo la hipótesis de MFD exacto, toda la información de las series $\{X_{jt}\}_{t=1}^{T}$ con $j \neq i$ relativa a la serie $\{X_{it}\}_{t=1}^{T}$ se recoge en los factores y en su propia componente idiosincrática. Esta es una de las principales motivaciones del MFD, ya que con un número reducido de series se sintetiza toda la información relevante presente en N series, con N habitualmente mucho más grande que r. Así, podemos estudiar la estructura y dinámica o realizar predicciones de forma más clara e interpretable con menos variables, algo que siempre interesa en econometría de cara a comprender el movimiento de las variables.

Modelo Factorial Dinámico aproximado Barhoumi *et al.* (2018) recogen una serie de limitaciones del MFD exacto que justifican considerar el modelo aproximado:

- El número de variables N suele ser mayor que el de observaciones, T, en las series económicas.
- Las hipótesis de i.i.d y de matriz de covarianzas diagonal de las componentes idiosincráticas será en general demasiado estricta para el contexto económico. Esto puede llevar a problemas de mala especificación del modelo y consiguientes errores en la estimación de los factores.
- La estimación por máxima verosimilitud generalmente es inabarcable para los MFD de alta dimensión al crecer demasiado el número de parámetros a estimar.

En la práctica es difícilmente justificable la penúltima condición ya que al considerar un número elevado de series cabe esperar que exista correlación entre algunas de ellas y por tanto entre sus componentes idiosincráticas. Bajo el modelo aproximado, la expresión (2.5) sería análoga pero contendría

términos extra reflejando la correlación entre perturbaciones. A la hora de realizar estimaciones se supondrá en general este modelo ya que es el que cabe esperar en la práctica. Para poder emplear este modelo se impondrán condiciones para que a pesar de permitirla, limitar la correlación entre las componentes idiosincráticas.

Giannone y Reichlin (2006) definen el MFD aproximado con las hipótesis:

$$0 < \liminf_{n\to\infty} \frac{1}{N}\lambda_{\min}(\lambda(L)'\lambda(L)) \leq \limsup_{n\to\infty} \frac{1}{N}\lambda_{\max}(\lambda(L)'\lambda(L)) < \infty$$

$$0 < \liminf_{n\to\infty} \frac{1}{N}\lambda_{\min}(\Sigma_e) \leq \limsup_{n\to\infty} \frac{1}{N}\lambda_{\max}(\Sigma_e) < \infty$$

siendo $\lambda_{\min}(\cdot)$ y $\lambda_{\max}(\cdot)$ el mínimo y máximo autovalor de la matriz respectivamente.

La primera hipótesis asegura que para N suficientemente grande, $\lambda(L)'\lambda(L)/N$ sea de rango completo r y por tanto que los factores afectan a todas las series. La segunda permite la correlación de las componentes idiosincráticas pero la limita. También garantiza que la varianza de las perturbaciones está acotada y es mayor que cero, lo que implica que no puede existir ninguna serie sin componente idiosincrática.

Una importante ventaja de considerar MFDs es que si los factores son conocidos y (ϵ_t, η_t) son gaussianos, se pueden realizar predicciones para una serie individual observada a partir de la regresión de la propia serie sobre sus retardos y los factores. Bajo el MFD exacto, $E[\epsilon_{it}\epsilon_{jt}] = 0$ si $i \neq j$. Así, se utiliza la ventaja de contar con la información de N variables condensada en solo r factores, donde $r << N$. Supongamos que $\{e_{it}\}_{t=1}^{T}$ sigue la estructura autoregresiva (2.4). Bajo el MFD exacto, y utilizando un criterio de mínimos cuadrados para la estimación, el estimador óptimo para la predicción a 1 paso de la i-ésima variable es:

$$\begin{aligned}
E[X_{i,t+1}|X_t, f_t, X_{t-1}, f_{t-1}, \dots] = \\
E[\lambda_i(L)f_{t+1} + e_{i,t+1}|X_t, f_t, X_{t-1}, f_{t-1}, \dots] = \\
E[\lambda_i(L)f_{t+1}|X_t, f_t, X_{t-1}, f_{t-1}, \dots] + E[e_{i,t+1}|X_t, f_t, X_{t-1}, f_{t-1}, \dots] = \\
E[\lambda_i(L)f_{t+1}|f_t, f_{t-1}, \dots] + E[e_{i,t+1}|e_{it}, e_{i,t-1}, \dots] = \\
\alpha_i(L)f_t + \delta_i(L)X_{it},
\end{aligned} \tag{2.5}$$

donde $\alpha_i(L) = \lambda_{i0}\Psi(L) - \delta_i(L)\lambda_i(L) + L^{-1}(\lambda_i(L) - \lambda_{i0})$. La tercera igualdad se sigue de (2.1), (2.2), (2.4) y la última de (1.2) y de la suposición de un MFD exacto. Nótese que con este modelo no aumenta el número de términos para el forecast al añadir más series al conjunto de observaciones del sistema. Si e_t y η_t están bajo las condiciones del MFD pero no son gaussianos, (2.5) se interpreta como un predictor lineal de la población. La ecuación (2.5) resume una propiedad clave de este modelo, y es que podemos explicar el movimiento de la serie y realizar predcciones una vez se tengan los factores, sin tener que considerar la información redundante que supondrían las demás series y el consiguiente

aumento del número de parámetros con los que lidiar. Se ha expresado solo la predicción a un paso, pero la definición anterior es fácilmente generalizable para cualquier otro horizonte h.

Cabe destacar que en caso de considerar (2.5) con $t = 1, \ldots, T$ en la ecuación no entrarían lags de la propia variable. Esto se debe a la suposición de que toda la información se recoge en los factores y que ninguna serie tiene poder predictor sobre otra, mostrando de nuevo la motivación para emplear este modelo,

$$E[X_{it}|X_t^{-i}, f_t, X_{t-1}^{-i}, f_{t-1}, \ldots] = \lambda_i(L)f_t$$

donde X_t^{-i} denota el vector de todas las variables menos la i-ésima. Vemos que la componente común para la i-ésima serie es la esperanza de X_{it} dados los factores y las demás variables, las cuales al no tener ningún poder explicativo sobre $\{X_{it}\}_{t=1}^T$ se "aglutinan" en los factores.

2.2. Forma estática del Modelo Factorial Dinámico

Sea $F_t = (f_t', f_{t-1}', \ldots, f_{t-p}')'$ el $(q \times 1)$ vector de factores estáticos. Sea también $\Lambda = (\lambda_0, \lambda_1, \ldots, \lambda_p)$, donde λ_h es la matriz de coeficientes $(N \times r)$ asociada al h-ésimo retardo en $\lambda(L)$. Definamos también Φ una matriz compuesta por $1s$, $0s$ y los elementos de $\Psi(L)$ tal que la estructura en (2.2) se reescriba en términos de F_t. Con esta notación se reescribe el MFD (2.1) y (2.2) en su forma estática como:

$$X_t = \Lambda F_t + e_t \tag{2.6}$$

$$F_t = \Phi F_{t-1} + G\eta_t, \tag{2.7}$$

donde $G = (I_r, 0_{r\times(q-r)})'$ es una matriz de $1s$ y $0s$ tal que (2.7) y (2.2) son equivalentes.

El término "estático" en esta representación se debe a que a pesar de que F_t es en este caso una serie de tiempo q-dimensional, al incorporar todos los retardos de f_t susceptibles de aparecer en el modelo en F_t, no es necesaria una interpretación como serie de tiempo de F_t para el ajuste del MFD. Por ejemplo, si se opta por la estimación por CP, los factores se estiman a partir de la descomposición por CP clásica, sin suponer ninguna dinámica o correlación serial. Esta forma consiste en reescribir el MFD (2.1) y (2.2) para que dependa de q factores estáticos F_t en lugar de r factores dinámicos, donde $r \leq q = (p+1)r$ y p es el grado de las matrices polinomiales introducidas en (2.1) y (2.2). La motivación primordial para esta representación radica en que el modelo estático es abordable desde un análisis de componentes principales y otro tipo de métodos por mínimos cuadrados, sin necesidad de seleccionar parámetros auxiliares.

Como ejemplo vamos a desarrollar el propuesto en Stock y Watson (2016). Consideremos un MFD con un solo factor dinámico ($r = 1$) donde $\{X_{it}\}_{t=1}^T$ dependa del valor actual y un retardo del factor

y que la dinámica VAR para $\{f_t\}_{t=1}^T$ en (2.2) tenga dos lags, es decir, $f_t = \Psi_1 f_{t-1} + \Psi_2 f_{t-2} + \eta_t$. La correspondencia entre el MFD dinámico y estático para la serie $\{X_{it}\}_{t=1}^T$ es:

$$X_{it} = \lambda_{i0} f_t + \lambda_{i1} f_{t-1} + e_{it} \;=\; (\lambda_{i0}\ \lambda_{i1}) \begin{pmatrix} f_t \\ \\ f_{t-1} \end{pmatrix} \;=\; \Lambda_i F_t + e_{it}$$

$$F_t = \begin{pmatrix} f_t \\ \\ f_{t-1} \end{pmatrix} = \begin{pmatrix} \Psi_1 & \Psi_2 \\ \\ 1 & 0 \end{pmatrix} \begin{pmatrix} f_{t-1} \\ \\ f_{t-2} \end{pmatrix} + \begin{pmatrix} I_r \\ \\ 0 \end{pmatrix} \eta_t \;=\; \Phi F_{t-1} + G\eta_t$$

donde $\Lambda_i = (\lambda_{i0}\ \lambda_{i1})$ es la i-ésima fila de Λ. Los dos desarrollos anteriores muestran la equivalencia entre los modelos y como se expresa un modelo respecto a su análogo. Puede generalizarse definiendo las matrices correspondientes de forma adecuada para cualquier MFD. Llegados a este punto donde el modelo dinámico y estático son equivalentes, el desarrollo del modelo según una representación u otra vendrá motivado por el problema en sí. Veremos más adelante que dependerá del número de series N la representación por la que se opte.

Al igual que en la forma dinámica del modelo, tenemos una importante reducción de la dimensionalidad de cara a definir y predecir la dinámica de las series a partir de los factores. Suponiendo que la componente idiosincrática sigue la autoregresión (2.4) y que las perturbaciones (ϵ_t, η_t) son gaussianas, en la versión estática del MFD la predicción a 1 paso (el análogo estático de (2.5)) es:

$$E[X_{i,t+1}|X_t, F_t, X_{t-1}, F_{t-1}, \dots] = \alpha_i(L)F_t + \delta_i(L)X_{it}, \tag{2.8}$$

donde $\alpha_i(L) = \Lambda_i\Phi - \delta_i(L)\Lambda_i$. Si las perturbaciones no son gaussianas, la expresión (2.8) se interpreta de nuevo como el predictor lineal de la población. Las predicciones a varios horizontes se pueden hacer a partir de una regresión en los valores pasados de $\{F_t\}_{t=1}^T$ y $\{X_{it}\}_{t=1}^T$ o iterando hacia delante a partir de (2.4) y (2.7).

2.2.1. Normalización de los factores

Debido a que los factores son inobservables solo son identificables bajo una serie de normalizaciones arbitrarias. En el MFD estático, el espacio abarcado por $\{F_t\}_{t=1}^T$ es identificable pero no los propios factores $\{F_t\}_{t=1}^T$. Sea Q una matriz invertible $(q \times q)$ cualquiera, entonces $\Lambda F_t = (\Lambda Q^{-1})(QF_t)$. A la hora de estudiar las dinámicas macroeconómicas o realizar predicciones, la falta de identificación no es un problema, ya que solo es necesario conocer el espacio generado por los factores. De esta forma, Q resulta irrelevante a la hora de estimar el modelo. En la práctica, al ser necesaria una normalización arbitraria, la falta de identificación se resuelve con algunas normalizaciones convenientes para facilitar

la estimación y que sea consistente. La matriz Q arbitraria considerada antes es $(q \times q)$ por lo que tiene q^2 parámetros libres y necesitaremos q^2 restricciones para la identificación del modelo.

Normalización por Componentes Principales Heredamos el nombre y la formulación en Bai y Ng (2008) que denotan así a esta normalización ya que se utiliza en la estimación por CP. Bajo esta normalización las columnas de Λ son ortogonales y tienen norma unitaria:

$$N^{-1}\Lambda'\Lambda = I_q \quad \text{y} \quad \Sigma_F \quad \text{diagonal} \tag{2.9}$$

donde $\Sigma_F = E[F_t F_t']$. La primera condición impone $q(q+1)/2$ restricciones y la segunda $q(q-1)/2$, lo que suma q^2 y completa la identificación del modelo. Esta normalización es muy intuitiva pero en desarrollos posteriores supondremos condiciones equivalentes por conveniencia para la estimación.

Normalización renombrando los factores Otra alternativa es asociar cada factor con una de las variables observables. Así lo que se hace es "renombrar" al factor con el nombre de la variable. La estandarización consiste en reordenar X_t tal que las primeras q variables son las que renombran a los factores. La normalización entonces es:

$$\Lambda = (I_q,\ \Lambda_{(q+1):n}), \quad \Sigma_F \text{ sin restricciones}$$

La componente común de $\{X_{it}\}_{t=1}^T$ es así $\{F_{it}\}_{t=1}^T$ para las q primeras variables y para las restantes una combinación de los factores. La submatriz $(q \times q)$ de cargas de las q primeras variables debe ser invertible. Esta condición es necesaria para que el espacio generado por las innovaciones de las q primeras componentes comunes abarquen el espacio generado por las innovaciones de los factores estáticos. En la práctica, las variables que renombran deben ser suficientemente diferentes y representativas de los distintos grupos de variables para que sea posible cubrir el espacio generado por los factores.

2.3. Modelo Factorial Dinámico estructural

Un objetivo recurrente en el marco económico ha sido estimar el efecto en la economía de perturbaciones estructurales, típicamente referidas como shocks. Ejemplos de shocks son un repentino salto en el precio del crudo, un inesperado cambio en la productividad o un aumento súbito de la demanda. Estos shocks introducen cambios inesperados de las variables económicas. Por definición, los shocks son autónomos e incorrelados con otros shocks. Además, al ser súbitos, son serialmente incorrelados.

Como es imposible observar las series temporales sin los shocks, se han desarrollado diversos métodos para identificarlos bajo las mínimas suposiciones extra. Fuera del marco de los MFDs, el enfoque predominante ha sido el VAR estructural (SVAR). La premisa de este modelo es que las innovaciones

de una serie de tiempo abarca los shocks. El problema de este enfoque es que al igual que en el modelo VAR, resulta inabordable la estimación cuando el número de observaciones es grande. En caso de considerar un número reducido de series de tiempo en la estructura del modelo VAR, estaríamos introduciendo un sesgo al no introducir información potencial para los shocks. Aumentar N solventaría este problema pero nos llevaría a una mayor complejidad computacional. Aquí, es donde el MFD puede ser utilizado para lidiar con este problema. El número de parámetros en el modelo VAR se incrementa con N^2 mientras que el MFD con N. El marco de los MFDs también facilita la medida de los errores, a partir de los factores, dotándonos de otra herramienta para identificar shocks.

El MFD estructural (MFDE) extiende directamente el modelo SVAR. En el MFDE todos los factores son inobservables. Con una pequeña modificación, uno o más de los factores pueden ser tratados como observables, convirtiendo el MFDE en un modelo VAR aumentado con factores (FAVAR por sus siglas en inglés). Más adelante profundizaremos más en el modelo FAVAR y en como incorporar la presencia de shocks.

Capítulo 3

Estimación del modelo

Una vez introducidas las dos formas del MFD, procedemos con la estimación de los factores, o más bien el espacio definido por estos al no ser observables. Existe una relación directa entre las dos representaciones del modelo, por lo que esencialmente, bajo cualquiera de los dos enfoques, lo que estaremos haciendo es estimar el espacio generado por los factores y su relación con las variables observadas, con la única diferencia de que bajo el enfoque dinámico se ajustan las relaciones de los lags de forma explícita y con el estático de forma implícita. Los primeros MFDs propuestos se consideraban en escenarios con una baja dimensión (N pequeño) y a partir de una representación del MFD como un modelo de espacio de estados se ajusta de forma paramétrica en el dominio temporal a partir del estimador de máxima verosimilitud (EMV) y el filtro de Kalman, el cual puede ser aplicado en este contexto al ser lineal el modelo en las variables no observables. Este método obtiene estimaciones óptimas para los factores f_t cuando nos encontramos bajo las suposiciones introducidas para este modelo. Sin embargo, requiere trabajar con optimizaciones no lineales, las cuales implican una restricción en el número de parámetros y por tanto del número de series N a considerar. Stock y Watson (1998) emplean técnicas paramétricas en su trabajo y Stock y Watson (1998a) motivan la aplicación de un enfoque no paramétrico, ya que entre otras cosas, posibilita considerar un numero de series N mayor. Como indican Sargent y Sims (1977), en vez de reducir la dimensionalidad del problema restringiendo el número de ecuaciones como en las técnicas previas, se imponen ciertas condiciones que simplifiquen el problema. Este enfoque se basa en la estimación por CP del espacio generado por los factores, la cual veremos que es consistente. Si N es lo suficientemente grande, se obtiene una estimación de los factores con una precisión adecuada para poder ser introducidos como variables en regresiones subsiguientes para obtener el resto de coeficientes desconocidos del MFD, como las cargas por ejemplo.

En la literatura más reciente han aparecido procedimientos alternativos para la estimación del MFD, basados en métodos Bayesianos como indican Stock y Watson (2016) por ejemplo, particularmente útiles en el caso donde el modelo contenga componentes no lineales o elementos no gaussianos. Al no ser el foco del trabajo y debido a su poco desarrollo no se introducen estas técnicas en esta memoria.

Ante paneles balanceados se suele optar por el enfoque no paramétrico, ya que el método de CP (o alguna técnica derivada como las que se verán a continuación) tienen numerosas ventajas. En la literatura actual, cuando N es pequeño, el MFD se expresa como un modelo de espacio de estados, se asume normalidad y los parámetros se estiman maximizando la función de verosimilitud gaussiana vía el filtro de Kalman. Ante paneles grandes, sobretodo al aumentar N, se recurre al análisis de CP, ya que el número de factores que se pueden estimar por este método es mín $\{N, T\}$, mucho mayor que los permitidos por el del espacio de estados. Esto resultará de interés a la hora de estimar los factores latentes cuando nos encontremos ante paneles grandes, donde cabe esperar que sean necesarios varios para obtener un buen ajuste del panel. Además, según indica Bai (2003), el estimador por CP de los factores converge al estimador de máxima verosimilitud cuando N crece. El modelo de espacio de estados tiene problemas computacionales también conforme aumenta N, llegando en algunos casos a ser imposible su implementación. Considerar la representación del MFD en su forma estática o dinámica viene motivado en gran parte por la técnica de estimación que se vaya a emplear. Si optamos por un enfoque paramétrico se interpreta el MFD como un modelo del estado de espacios, donde la representación dinámica es la apropiada. En caso de escoger el planteamiento no paramétrico, el ajuste del modelo se calcula por medio de CP o alguna generalización, por lo que se considera la forma estática del MFD.

En los software estadísticos como R, los (pocos) paquetes desarrollados para la modelización de MFDs se inclinan por el uso del algoritmo de Esperanza-Maximización (EM), introducido más adelante. Esto se debe a que generaliza el análisis por CP con un algoritmo iterativo que puede ser empleado ante paneles desbalanceados, lo que permite explotar las ventajas de estimar un modelo de forma no paramétrica a pesar de los datos ausentes. Como se comentó anteriormente, en la literatura se sugiere el uso del análisis por CP, ya que en general es difícil justificar la elección de los parámetros en caso de optar por un enfoque paramétrico y permite estimar un mayor número de factores. La tendencia general a emplear el algoritmo EM también se deba probablemente a que está más desarrollado y estudiado en otros campos, lo que ha facilitado la implementación a partir de trabajos previos. El otro algoritmo iterativo disponible en el entorno de R es el de *dos pasos*, también introducido más adelante. Este enfoque resulta muy intuitivo, pero computacionalmente es menos eficiente al requerir de más estimaciones intermedias y ser exclusivo al ajuste de MFDs, por lo que no ha sido tan desarrollado como el EM.

3.1. Métodos no paramétricos

A parte de lo introducido hasta ahora, la motivación para considerar este enfoque radica en que en una media ponderada desde la sección cruzada de X_t, $\bar{X}_t = N^{-1} \sum_{i=1}^{N} X_{it}$, las perturbaciones idiosincráticas tenderán a cero por la ley débil de los grandes números, por lo que solo permanecen en la expresión las combinaciones lineales de los factores y por tanto solo las variaciones asociadas a estos. Los métodos no paramétricos obtienen estimadores de los factores estáticos en (2.6) sin necesidad de

especificar o asumir ningún tipo de modelo o distribución para las perturbaciones. Tampoco se impone ningún parámetro para las dinámicas idiosincráticas como en (2.4). En lugar de ello, F_t se considera como un parámetro q-dimensional a estimar a partir del vector de datos N-dimensional X_t. Se hacen suposiciones más débiles que las paramétricas, basadas en el MFD aproximado y en la estructura de los factores, en vez de sobre los parámetros en sí. Se introducen debido a que las cargas y los factores no son separables y en aras de obtener un modelo consistente. En la práctica, cabe esperar que el MFD no sea exacto, por lo que las hipótesis introducidas son importantes en aras de no obtener un modelo erróneo. Estas hipótesis se satisfacen bajo el MFD exacto, por lo que no suponen ninguna pérdida de generalidad.

Mostremos un ejemplo para motivar la utilización de este tipo de técnicas no paramétricas. Supongamos un MFD con un solo factor. En ese caso, la media en la sección cruzada de X_t se expresa como $\bar{X}_t = \bar{\Lambda} F_t + \bar{e}_t$, siendo $\bar{\Lambda}$ y $\bar{e}_t$ las correspondientes medias en la sección cruzada. Si la correlación en la sección cruzada de $\{e_{it}\}$ está limitada, $\bar{e}_t \xrightarrow{\mathbb{P}} 0$ (ya que las componentes idiosincráticas se suponen con media cero), y por tanto $\bar{X}_t - \bar{\Lambda} F_t \xrightarrow{\mathbb{P}} 0$. Por tanto, si $\bar{\Lambda} \neq 0$, $\bar{X}_t$ estima los factores estáticos F_t salvo escala. El argumento anterior puede ser generalizado para el caso en que el modelo está conformado por más de un factor considerando una media ponderada de X_t.

3.1.1. Componentes Principales

En esta sección se asumen las siguientes hipótesis para motivar la técnica de estimación a partir del MFD aproximado. Más adelante se amplían y se matizan para poder ajustar el modelo bajo condiciones más generales. Estas condiciones son necesarias para la identificación del modelo, en cuanto a que los factores y las cargas no son separadamente identificables:

$$N^{-1}\Lambda'\Lambda \overset{N\to\infty}{\longrightarrow} D_\Lambda \tag{3.1}$$

$$\text{máx eigen}(\Sigma_e) \leq c < \infty \ \ \forall N, \tag{3.2}$$

donde la matriz de covarianzas D_Λ $(q \times q)$ es de rango completo, máx eigen denota el operador máximo autovalor y $\Sigma_e = E[e_t e_t']$. Notemos que con estas restricciones se identifica el modelo ya que la primera hipótesis introduce $q(q-1)$ restricciones y la segunda q. Además de identificar el modelo, nos sitúan en un marco donde obtenemos estimadores de los parámetros del modelo convenientes para la posterior interpretación y facilitar sucesivos cálculos. La condición (3.1) asegura que el efecto de los factores afecta a la mayor parte de las series y que las cargas son heterogéneas (las columnas de Λ no son muy similares). Esto resulta conveniente en el sentido de que entre todas las posibles soluciones, imponemos obtener una donde las cargas asociadas a cada serie sean lo más distintas dos a dos. Resulta interesante que una serie tenga una carga mayor en un factor y otra serie en otro para interpretar el modelo y las dinámicas de las series, así como estudiar qué series tienen más peso en la construcción de cada factor. Ambas características son deseables en cuanto a que de esta forma estamos estimando factores que influyen y explican significativamente las series consideradas (todas las variables están relacionadas

con los factores estimados) y que las cargas nos permiten separar e identificar las series con un elevado grado de claridad al buscar una estimación donde las cargas sean lo menos parecidas dos a dos. Si todas las cargas fuesen similares obtendríamos una representación muy parecida de todas las series, lo cual no resultaría de interés ni permitiría estudiar qué variables influyen más y en qué grado al movimiento de los factores. La condición (3.2) limita la correlación entre las componentes idiosincráticas de las series. Ya comentamos que en la estimación no paramétrica nos basaremos en el modelo aproximado, donde se permite un cierto grado de correlación entre las perturbaciones, pero debemos limitarla para que el modelo sea abordable.

En Stock y Watson (2006) se muestra que si Λ fuese conocida, F_t se podría estimar como $(\Lambda'\Lambda)^{-1}\Lambda' X_t$, expresión que coincide con el estimador de los parámetros en el modelo lineal general. Por (3.1) y (2.6), $(\Lambda'\Lambda)^{-1}\Lambda' X_t = F_t + (\Lambda'\Lambda)^{-1}\Lambda' e_t$, así que bajo la hipótesis de varianza limitada de los factores, F_t podría estimarse con precisión si N es lo suficientemente grande.

Como primera intuición en el funcionamiento de esta media en la sección cruzada, consideremos un estimador de F_t a partir de la suma ponderada de X_t con la matriz (no aleatoria) $W \in \mathcal{M}_{N\times q}$. Consideramos que la matriz W está normalizada tal que $W'W/N = I_q$, donde I_q es la matriz identidad de orden q. El estimador de esta forma sería:

$$\hat{F}_t(N^{-1}W) \;=\; N^{-1}W'X_t \tag{3.3}$$

Si $N^{-1}W\Lambda \to H$ según $N \to \infty$, con H una matriz $(q \times q)$ de rango completo y si se cumple las condiciones (3.1) y (3.2), entonces $\hat{F}_t(N^{-1}W)$ es un estimador consistente para el espacio abarcado por F_t:

$$\hat{F}_t(N^{-1}W) \;=\; N^{-1}W'(\Lambda F_t + e_t) \;=\; N^{-1}W'\Lambda F_t + N^{-1}W'e_t \;\xrightarrow{\mathbb{P}}\; HF_t, \quad \textit{cuando } N \to \infty \tag{3.4}$$

ya que se supone $N^{-1}W'\Lambda \to H$ y que $N^{-1}W'e_t \xrightarrow{\mathbb{P}} 0$ por la ley débil de los grandes números [1]. Al ser H de rango completo, el estimador $\hat{F}_t(N^{-1}W)$ estima consistentemente el espacio abarcado por los factores. La clave para garantizar esto es tener en el desarrollo (3.4) una matriz de pesos tal que $N^{-1}W'\Lambda \to H$.

El objetivo principal es aproximar los factores por medio de una combinación lineal de los datos tal que se maximice la varianza de los factores estimados, $\Sigma_F = W'\hat{\Sigma}_X W$, donde $\hat{\Sigma}_X = (1/T)\sum_{t=1}^{T} X_T X_t'$ es la matriz de covarianzas muestral del vector de datos X_t.

El análisis de CP entra en juego para la estimación de la matriz de pesos. En general, no habrá la suficiente estructura en Λ como para proponer una matriz de pesos W que no dependa de los datos, X_t. Aquí es donde aparece el análisis de CP. En esta sección seguiremos el enfoque de Stock y Watson (2011). El estimador por CP de F_t es la suma ponderada definida en (3.3), con $W = \hat{\Lambda}$, donde $\hat{\Lambda}$ es la

[1] Esto se sigue de la desigualdad de Chebyschev. Sea W_j la j-ésima columna de W. Entonces $\mathrm{Var}(N^{-1}W_j'e_t) \leq$ máx eigen$(\Sigma_e)/N \leq c/N \to 0$. La primera desigualdad se sigue de que W está normalizada y la segunda de (3.2)

matriz de autovectores asociada a los q mayores autovalores de $\hat{\Sigma}_X$. Como muestran Bai y Ng (2002), $\hat{\Sigma}_X$ es un estimador consistente de Σ_X.

El estimador por CP se obtiene resolviendo el problema de mínimos cuadrados:

$$\min_{F_1,\ldots,F_T,\Lambda} V_q(\Lambda, F),$$

donde

$$V_q(\Lambda, F) \;=\; \frac{1}{NT}\sum_{t=1}^{T}(X_t - \Lambda F_t)'(X_t - \Lambda F_t), \tag{3.5}$$

sujeto a la normalización $N^{-1}\Lambda'\Lambda = I_q$. En Stock y Watson (2016) se demuestra que bajo el MFD exacto, con varianza homogénea de las componentes idiosincrátcias y con los factores considerados como parámetros, la solución al problema de minimización (3.5) es el estimador de máxima verosimilitud gaussiana. Cabe destacar que esto es análogo a lo que ocurre en el modelo de regresión.

Para resolver (3.5) primero se minimiza F_t dado Λ y se obtiene $\hat{F}_t\left(\Lambda(\Lambda'\Lambda)^{-1}\right) \;=\; (\Lambda'\Lambda)^{-1}\Lambda' X_t$. Introduciendo esto en (3.5) el problema se convierte ahora en $\min_\Lambda T^{-1}\sum_{t=1}^{T} X_t'[I - \Lambda(\Lambda'\Lambda)^{-1}\Lambda']X_t$. Esta minimización es equivalente al problema $\max_\Lambda \operatorname{tr}\left\{(\Lambda'\Lambda)^{-1/2'}\Lambda'\left(T^{-1}\sum_{t=1}^{T} X_t X_t'\right)\Lambda(\Lambda'\Lambda)^{-1/2}\right\}$, que equivale a $\max_\Lambda \Lambda'\hat{\Sigma}_X\Lambda$ sujeto a $N^{-1}\Lambda'\Lambda = I_q$. Esto conforma el clásico problema de componentes principales, cuya solución consiste en establecer $\hat{\Lambda}$ igual a los autovectores escalados de $\hat{\Sigma}_X$ correspondientes a sus q mayores autovalores. Se concluye que el estimador de mínimos cuadrados de F_t es $\hat{F}_t \;=\; N^{-1}\hat{\Lambda}' X_t$, que coincide con las primeras q componentes principales escaladas de X_t.

En Stock y Watson (1998) se muestra que la estimación por CP permanece consistente cuando hay pequeñas variaciones temporales en Λ o leve contaminación en los datos si disponemos de un número de predictores $N >> T$. Bai (2003) muestran que el estimador por CP es asintóticamente equivalente al de máxima verosimilitud si se asume normalidad.

3.1.2. Consistencia de la estimación de los factores y ecuación de predicción

Vamos a recoger una serie de resultados para mostrar la consistencia del método. Introducidos en Stock y Watson (2002a), son asintóticos en cuanto a que se supone que $N, T \to \infty$ conjuntamente. La suposición de estas hipótesis aseguran la identificación del modelo y su correcta estimación.

Suposiciones FC Se hacen sobre los factores y las cargas,

FC1. $(\Lambda'\Lambda/N) \to I_q$

FC2. $E[F_t F_t'] = \Sigma_F$, donde Σ_F es una matriz diagonal con elementos $\sigma_{ii} > \sigma_{jj} > 0$ para $i < j$.

FC3. $|\lambda_{ij}| \leq \bar{\lambda} < \infty, \forall i \in \{1, \cdots, N\}, j \in \{1, \cdots, q\}$

FC4. $T^{-1} \sum_{t=1}^{T} F_t F_t' \xrightarrow{\mathbb{P}} \Sigma_F$

donde λ_{ij} es el elemento (i, j) de la matriz Λ. Estas hipótesis sirven para identificar los factores. Las dos primeras implican que los factores proporcionan una contribución no despreciable a la varianza de $\{X_{it}\}_{t=1}^{T}$. También se solventa el problema de identificación como en (2.9). Como se comentó en la Sección 2.2.1, para cualquier matriz invertible Q, $\Lambda F_t = \Lambda Q Q^{-1} F_t$, lo que implica tener que suponer algún tipo de restricción para la identificación de los factores y de las cargas separadamente. FC1 restringe Q a ser ortonormal y FC2 a que sea diagonal. De esta forma Q se considera una matriz diagonal con elementos ± 1 por lo que la identificación está resuelta salvo por un término ± 1, es decir, se identifican los factores salvo por un cambio de signo.

Equivalentemente, FC proporciona la normalización anterior (asintóticamente) asociando Λ con los autovectores ordenados de $(NT)^{-1} \sum_{t=1}^{T} \Lambda F_t F_t' \Lambda'$ y F_t con las CP de ΛF_t. Los elementos de la diagonal de Σ_F se corresponden al límite de los autovalores de $(NT)^{-1} \sum_{t=1}^{T} \Lambda F_t F_t' \Lambda'$, que se suponen distintos por conveniencia. En caso de no ser distintos, los factores solo podrían identificarse tras una transformación ortonormal. La condición FC2 permite que los factores sean serialmente correlados, que se satisface por ejemplo si los factores dinámicos f_t son estacionarios en covarianza.

Suposiciones M Sobre los momentos de los errores e_t,

M1. $E[e_t' e_{t+s}/N] = \gamma_{N,t}(s)$ y $\lim_{N\to\infty} \sup_t N^{-1} |\gamma_{N,t}(s)| < \infty$.

M2. $E[e_{it} e_{jt}] = \tau_{ij,t}$ y $\lim_{N\to\infty} \sup_t N^{-1} \sum_{i=1}^{N} \sum_{j=1}^{N} |\tau_{ij,t}| < \infty$

M3. $\lim_{N\to\infty} \sup_{t,s} N^{-1} \sum_{i=1}^{N} \sum_{j=1}^{N} |\mathrm{Cov}(e_{is} e_{it}, e_{js} e_{jt})| < \infty$

La primera hipótesis permite correlación temporal en e_t y la segunda correlación débil entre las series. La normalidad no se asume, pero se limita el tamaño del momento de orden 4 con M3.

Suposiciones Y Sobre la ecuación de predicción (1.2). Sea $z_t = (F_t', w_t')'$ y $\beta = (\alpha, \delta)$. Entonces:

Y1. $E[z_t z_t'] = \Sigma_z$ es una matriz definida positiva

Y2. $T^{-1} \sum_{t=1}^{T} z_t z_t' \xrightarrow{\mathbb{P}} \Sigma_z$

Y3. $T^{-1} \sum_{t=1}^{T} z_t \varepsilon_{t+h} \xrightarrow{\mathbb{P}} 0$

Y4. $T^{-1} \sum_{t=1}^{T} \varepsilon_{t+h}^2 \xrightarrow{\mathbb{P}} 0$

Y5. $|\beta| < \infty$

Recordemos que en (1.2) w_t denota un un vector $m \times 1$ de variables observables, por ejemplo lags de Y_t o alguna variable que pueda tener un elevado poder predictor sobre Y_t conjuntamente con f_t. Los parámetros α y δ son respectivamente los coeficientes de la regresión de Y_t sobre los factores y w_t, y ε_t representa el error en el ajuste.

Las tres primeras aseguran la consistencia de la regresión (1.2) y las restantes que se conserve la consistencia cuando el forecast se realiza con los factores estimados y no con los verdaderos.

El siguiente resultado muestra que el estimador por CP es puntualmente (para cualquier t) consistente, y con error cuadrático medio (MSE) limitado. Esto será de suma importancia a fin de poder considerar correctamente la estimación de los factores en lugar de los verdaderos en sucesivos cálculos y/o regresiones. Como la suposición FC no identifica el signo de los factores, el siguiente teorema se formula en términos de los factores estimados y ajustados para tener el signo correcto.

Teorema 1. *(Teorema 1 de Stock y Watson (2002a)). Sea S_i una variable con valor ± 1, $N, T \to \infty$ y supongamos que se satisfacen FC y M. Supongamos que se estiman k factores, siendo q el número verdadero de factores latentes. Entonces, se puede seleccionar S_i tal que se cumpla:*

- *Para* $i = 1, \ldots, q$, $\quad T^{-1} \sum_{t=1}^{T} (S_i \hat{F}_{it} - F_{it})^2 \xrightarrow{\mathbb{P}} 0$
- *Para* $i = 1, \ldots, q$, $\quad S_i \hat{F}_{it} \xrightarrow{\mathbb{P}} F_{it}$
- *Para* $i = q+1, \ldots, k$, $\quad T^{-1} \sum_{t=1}^{T} \hat{F}_{it}^2 \xrightarrow{\mathbb{P}} 0$

Notemos que no se ha supuesto que $k = q$. La demostración del teorema se recoge en la demostración del Teorema 1 en Stock y Watson (2002a). La prueba sería sencilla en caso de conocer la matriz de cargas Λ, ya que F_t podría ser estimado con una regresión por mínimos cuadrados de $\{x_{it}\}_{t=1}^{T}$ sobre $\{\Lambda_i\}_{i=1}^{N}$. La consistencia del estimador se estudia analizando $\hat{F}_t - F_t = (\Lambda'\Lambda/N)^{-1}(N^{-1} \sum_i \Lambda_i e_{it})$. Por las hipótesis $N \to \infty$, $(\Lambda'\Lambda/N) \to I_q$ (FC1) y $N^{-1} \sum_i \Lambda_i e_{it} \xrightarrow{\mathbb{P}} 0$ (M1 y FC3), se sigue la consistencia de $\hat{F}_t$.

Análogamente, si se conociesen los verdaderos factores F_t, se podría estimar Λ_i con una regresión de $\{x_{it}\}_{i=1}^{N}$ sobre $\{F_t\}_{t=1}^{T}$. Se llegaría de nuevo a la consistencia analizando $(T^{-1} \sum_t F_t F_t')^{-1} T^{-1} \sum_t F_t e_{it}$ conforme $T \to \infty$ a partir de las hipótesis M1, FC2 y FC4.

En caso de no conocer ni los factores ni sus cargas, en el que nos encontramos en la práctica, se necesita que ambos índices $N, T \to \infty$. La demostración en este caso se basa en demostrar que los primeros q autovectores de Σ_X tienen la misma conducta que los de $T^{-1} \sum_{t=1}^{T} (\Lambda F_t)(\Lambda F_t)'$ (hipótesis M es crítica para esto) y después comprobar que estos autovectores pueden generar un estimador consistente para F_t (donde la hipótesis FC entra en juego).

El siguiente teorema supone que la predicción se estima utilizando $k = q$ factores. Esto no supone una pérdida importante de generalidad, ya que existen diversos métodos para estimar el número de factores de forma consistente. Estos se introducen en el siguiente capítulo.

Teorema 2. *(Teorema 2 de Stock y Watson (2002a)). Supongamos que se cumple Y y las condiciones del Teorema 1. Sean $\hat{\alpha}$ y $\hat{\delta}$ los estimadores por mínimos cuadrados en la regresión* (1.2)*. Entonces:*

- $(\hat{\alpha}'\hat{F}_T + \hat{\delta}w_T) - (\alpha' F_T + \delta w_T) \xrightarrow{\mathbb{P}} 0$
- $\hat{\delta} - \delta \xrightarrow{\mathbb{P}} 0$ *y S_i (análogo al del Teorema 1) se puede seleccionar tal que $S_i\hat{\alpha}_i - \alpha_i \xrightarrow{\mathbb{P}} 0$ para $i = 1, \ldots, q$.*

La demostración del Teorema 2 se sigue directamente del Teorema 1 con las hipótesis Y. Con este resultado tenemos garantizado que se puede estimar de forma consistente el ajuste de cualquier serie de interés respecto a los factores. Esto nos sitúa en un escenario en donde cualquier serie puede ser representada en función de los factores, lo que facilita su interpretación y sobre todo sus predicciones. Teniendo en cuenta además que las series de interés son susceptibles de no estar actualizadas en el momento de la predicción pero que los factores son estimados considerando toda la información disponible, de esta forma se tienen predicciones empleando más información y más instantes temporales que si no se considerasen los factores.

3.1.3. Componentes principales generalizadas

Este estimador es a las CP lo mismo que la regresión lineal generalizada a la regresión lineal simple. Si las perturbaciones son heterocedásticas o tienen algún tipo de correlación cruzada (la matriz Σ_e de covarianzas de e_t no es proporcional a la identidad), se puede mejorar la eficiencia de las estimaciones modificando la función (3.5) para dar cabida a una matriz de pesos más general. La regresión por mínimos cuadrados generalizada sugiere una versión ponderada de (3.5) por Σ_e^{-1}:

$$\min_{F_1,\ldots,F_T,\Lambda} T^{-1} \sum_{t=1}^{T} (X_t - \Lambda F_t)' \Sigma_e^{-1} (X_t - \Lambda F_t), \tag{3.6}$$

Una solución para el problema (3.6) es el estimador por CP, $\tilde{F}_t = N^{-1}\tilde{\Lambda}X_t$, donde $\tilde{\Lambda}$ son los autovectores escalados asociados a los q mayores autovalores de $\Sigma_e^{-1/2}\hat{\Sigma_X}\Sigma_e^{-1/2'}$. La matriz Σ_e es inobservable, por lo que se requiere su estimación. Se sustituye Σ_e por un estimador $\hat{\Sigma}_e$, el cual si es consistente concluye que la estimación por CP generalizadas es asintóticamente más eficiente que por CP.

Para cumplir la condición de consistencia, se han propuesto diversos estimadores para Σ_e. Boivin y Ng (2006) proponen una aproximación en dos pasos, donde primero se estiman los coeficientes

por CP y después se define como $\hat{\Sigma}_e$ una matriz diagonal cuyos elementos son la varianza muestral de los errores de la regresión de x_{it} sobre las CP estimadas en el primer paso. Forni *et al.* (2005) sugieren una generalización de la Identidad Fundamental del Análisis Factorial clásico, considerando la descomposición $\Sigma_X = \Sigma_{\Lambda F} + \Sigma_e = \Lambda \Sigma_X \Lambda' + \Sigma_e$, donde $\Sigma_{\Lambda F}$ es la matriz de covarianzas de la componente común ΛF_t. Esta descomposición se sigue de (2.6) y de que las componentes comunes y las perturbaciones sean incorreladas. Dada una estimación inicial por CP, proponen la aproximación $\hat{\Sigma}_e = \hat{\Sigma}_X - \hat{\Sigma}_{\Lambda F}$.

3.2. Métodos paramétricos

Estas técnicas se apoyan en la representación del MFD como un modelo del espacio de estados. Para ello es necesario especificar un modelo paramétrico para X_t, e_t y f_t en la versión dinámica del MFD. También son necesarias ciertas suposiciones sobre la distribución del vector de error y de las componentes idiosincráticas. En este enfoque se utiliza el filtro de Kalman para calcular la función de verosimilitud gaussiana y el filtro y suavizador de Kalman para obtener estimadores eficientes de los factores dada la muestra X_t. El primer paso es especificar el MFD como un modelo de espacio de estados para poder aplicar el filtro de Kalman.

Modelo de espacio de estados con factores dinámicos Esta es la primera representación que aparece en la literatura por su equivalencia directa con el espacio de estados. A partir de (2.1) y (2.2) el MFD estático se completa especificando una estructura para el proceso e_t y el término de error η_t. Habitualmente se supone una dinámica autorregresiva univariante en e_t. Expresamos (2.4) de forma compacta por conveniencia:

$$d_i(L)e_t = \xi_{it}, \quad i = 1, \ldots, N \tag{3.7}$$

donde $\xi_{it} \sim N(0, \sigma^2_{\xi_i})$ i.i.d con $i = 1, \ldots, N$, $\eta_{jt} \sim N(0, \sigma^2_{\eta_j})$ i.i.d con $j = 1, \ldots, r$ y $\{\xi_t\}$ y $\{\eta_t\}$ son procesos independientes. Las ecuaciones (2.1), (2.2), (3.7) con estas hipótesis de normalidad constituyen un modelo lineal de espacio de estados completo. En terminología del modelo de espacio de estados, (2.1) es la *ecuacion de observación* y (2.2) la *ecuación de transición*. A priori para poder expresar el MFD de esta forma es necesario que el panel de datos esté completo. Esto puede solventarse considerando solo un subconjunto de este o sustituyendo los valores ausentes por alguna estimación como la media, mediana o ceros. Como ya comentamos, en el MFD podremos solventar este problema vía el filtro de Kalman, por lo que no será un impedimento.

Modelo de espacio de estados con factores estáticos El modelo dinámico como se ha visto puede ser expresado de forma estática como (2.6), (2.7). La literatura reciente se inclina más por esta representación, ya que se puede resolver por máxima verosimilitud con el filtro de Kalman análogamente

pero permite mayor flexibilidad para lidiar con datos ausentes. Dados todos los parámetros del modelo, se aplica el filtro de Kalman para calcular la función de verosimilitud y para estimar los valores filtrados para F_t (y por tanto de f_t). Dada la estimación de los factores $\hat{F}_t$, Λ se estima a partir de una regresión de $\{x_t\}_{t=1}^{T}$ sobre $\left\{\hat{F}_t\right\}_{t=1}^{T}$ y los residuos de la regresión se emplean para estimar la autoregresión en (3.7). Los coeficientes de la representación en (2.7) se estiman a partir de $\hat{F}_t$ y la varianza de $G\eta_t$ se puede estimar a partir de los residuos del modelo VAR ajustado.

La principal ventaja de este método es poder trabajar con paneles que presenten irregularidades, como el problema de que algunas series se observen en frecuencias temporales heterogéneas (semanales, mensuales, ...). Si algunas series no están disponibles en algún instante se puede cambiar la dimensión de la ecuación de medida conforme los datos se van obteniendo. El problema es que el número de parámetros a estimar es proporcional a N, aumentando la dimensión de la ecuación de medida en la expresión (2.6).

En este tipo de modelos también se podría incluir la estimación a partir del algoritmo EM. Debido a que su aplicación principal radica en el escenario con datos faltantes y/o panel desbalanceado, se incluye en la sección específica a tal escenario. Cuando estemos ante un panel completo, las técnicas vistas hasta ahora aportan buenas aproximaciones como hemos visto, pero cuando se cuenta con datos ausentes recurriremos a técnicas de optimización derivadas a partir de las anteriores. Barhoumi *et al.* (2018) indican cómo la estimación por máxima verosimilitud puede llevar a problemas computacionales conforme aumenta el número de series, donde sugieren algún tipo de optimización numérica como el algoritmo EM.

3.2.1. Filtro de Kalman

Una vez representado el MFD estático (2.6), (2.7) como un modelo de espacio de estados se puede aplicar el filtro de Kalman de forma recursiva para obtener estimaciones eficientes. Para poder aplicarlo, debemos suponer que los errores idiosincráticos y los shocks de los factores siguen una distribución normal (la ecuación para η_t simplemente reescribe (2.3)):

$$\begin{aligned} e_t &\sim N(0, \Sigma_e), \quad e_t \in \mathbb{R}^N, \quad t = 1, \cdots, T \\ \eta_t &\sim N(0, \Sigma_\eta), \quad \eta_t \in \mathbb{R}^q, \quad t = 1, \quad \cdots, T \\ \xi_t &\sim N(0, \Sigma_\xi), \quad \xi_t \in \mathbb{R}^N, \quad t = 1, \cdots, T \end{aligned}$$

Aquí es donde aparece explícitamente el carácter paramétrico de esta técnica de estimación y uno de sus mayores problemas, ya que en la práctica no suele ser fácil suponer alguna estructura o parámetros para los errores.

Se muestra el filtro de Kalman en línea con el introducido en Arouba *et al.* (2007). Sea $\chi_t \equiv \{X_1, \cdots, X_t\}$, $a_{t|t} \equiv E[F_t|\chi_t]$, $P_{t|t} = Var(F_t|\chi_t)$ y $P_{t+1|t} = Var(F_{t+1}|\chi_t)$. A partir de unos valores

iniciales $a_{0|0}$ y $P_{0|0}$, la actualización del filtro de Kalman y la ecuación de predicción son:

$$\begin{aligned} a_{t|t} &= a_t + P_t\Lambda' S_t^{-1} v_t \\ P_{t|t} &= P_t - P_t\Lambda' S_t^{-1}\Lambda P_t' \\ a_{t+1|t} &= \Phi a_{t|t} \\ P_{t+1|t} &= \Phi P_{t|t}\Phi' + G\Sigma_\eta G' \end{aligned}$$

para $t = 1, \cdots, T$, donde $S_t = (\Lambda P_t \Lambda' + \Sigma_e)$ se corresponde con la varianza de la muestra X_t bajo el modelo y $v_t = X_t - \Lambda a_t$ el error del ajuste.

Notemos que a_{t+1} es la estimación del vector de estados para $t+1$ condicionada en la información disponible hasta t y P_{t+1} la correspondiente matriz de covarianzas. Pueden usarse como indican Camacho y Doménech (2012) para evaluar la función de log-verosimilitud:

$$l_t = -\frac{1}{2}[\ln(2\pi|F_{t|t}|) + v'_{t+1|t}(F_{t|t})^{-1} v_{t+1|t}],$$

donde $F_{t+1|t} = \Lambda P_{t|t}\Lambda'$ es la matriz de covarianzas del vector de estados en $t+1$ condicionada a la información disponible hasta t.

Una de las propiedades más primordiales del filtro de Kalman es que permanece válido con datos ausentes como indican Angelini *et al.* (2008). Si alguno de los elementos de X_t están ausentes, se remplaza la ecuación de medida con:

$$\begin{aligned} \tilde{X}_t &= \tilde{\Phi} F_t + \tilde{e}_t \\ \tilde{e}_t &\sim N(0, \tilde{\Sigma_e}) \end{aligned}$$

donde $\tilde{X}_t$ es de dimensión $\tilde{N} < N$ contiene los elementos de X_t que son observados. La clave para que tenga sentido considerar esto consiste en que X_t está relacionado con $\tilde{X}_t$ a través de la transformación $\tilde{X}_t = A_t X_t$, con $A \in \mathcal{M}_{N\times N}$. Nótese que en caso de que todos los datos de X_t estuvieran ausentes, la actualización de los estados se obviaría en el algoritmo y el filtro de Kalman proporcionaría un modelo con un forecast para todos los datos de todas las series.

3.3. Estimación con datos ausentes y frecuencia heterogénea

Panel desbalanceado Como ya se ha comentado, la ausencia de datos puede deberse a múltiples razones como que algunas series tienen una fecha de publicación más tardía que otras o el estar considerando variables medidas con distinta frecuencia. Cuando consideramos un número grande de series, es razonable considerar que nos podamos encontrar con datos faltantes para varias series. Dependiendo

de si estamos ante un enfoque paramétrico o no varía el tratamiento que haremos de los datos faltantes. En la literatura, todos los procedimientos para lidiar con esta situación suponen que las posiciones de los datos faltantes es aleatoria, es decir, que un dato esté ausente no depende de la propia variable. En el entorno económico esta suposición está justificada.

Una vez obtenidas las estimaciones de Λ y F_t para el escenario con datos ausentes a partir de los métodos adaptados para este contexto, se definen análogamente la componente común de la serie i-ésima como $\Lambda_i F_t$ y la predicción como en (2.8). Solo varía el desarrollo necesario para obtener los estimadores. Se adaptan las técnicas introducidas en las secciones anteriores para ajustar el modelo cuando nos encontramos ante un panel desbalanceado, pero una vez se tienen los estimadores para los factores y las cargas se interpretan y emplean como se ha introducido previamente.

Puede ser conveniente considerar un panel balanceado más grande que con el que contamos. En este contexto se puede lidiar con el panel incompleto permitiendo que la ecuación de medida (2.1) varíe dependiendo de si el dato está disponible o no en el instante t. Otra alternativa es introducir un valor proxy en los datos vacíos (la mediana de los datos o una media de los valores más cercanos temporalmente por ejemplo) y ajustar los parámetros del modelo para que estas observaciones no tengan peso a la hora de aplicar el filtro de Kalman.

Componentes Principales con datos ausentes La solución al problema de mínimos cuadrados (3.5) se sostiene cuando todas las NT observaciones están disponibles, i.e. la situación ante un panel balanceado. Cuando tratamos con datos faltantes se puede seguir utilizando el problema de mínimos cuadrados para estimar F_t y Λ, sin embargo, la solución debe obtenerse numéricamente. La modificación de (3.5) ante datos ausentes es:

$$\min_{F_1,\ldots,F_T,\Lambda} \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} S_{it}(x_{it} - \Lambda_i F_t)'(x_{it} - \Lambda_i F_t), \tag{3.8}$$

donde $S_{it} = 1$ si la observación x_{it} está disponible y $S_{it} = 0$ en caso contrario. El problema (3.8) se puede resolver iterativamente minimizando Λ dado F_t y después F_t dado Λ. Los valores iniciales suelen tomarse como los estimadores por CP dado un subconjunto de la muestra en donde no haya datos ausentes, por ejemplo el subpanel longitudinal en la Figura 1.1. Otra alternativa, propuesto por Stock y Watson (2002), es aplicar el algoritmo EM.

3.3.1. Estimación con el algoritmo de Esperanza-Maximización

A pesar de que el panel está desbalanceado, se puede obtener el estimador por mínimos cuadrados de F_t a partir de la función objetivo (3.8). Los métodos EM son ampliamente utilizados como algoritmos que extienden la estimación por máxima verosimilitud a los casos donde existen variables aleatorias inobservables en un modelo. En este contexto se trata de una modificación del análisis de CP para que

pueda ser empleado ante un panel desbalanceado, por tanto se considera la representación del MFD estático. Como indican Reis y Watson (2010), a pesar de la complejidad del modelo, la estructura lineal de las variables latentes hace que se pueda someter al algoritmo EM, con el paso "E" calculado con el filtro de Kalman y el "M" por una regresión lineal.

Sea $\theta = (F, \Lambda)$. El algoritmo EM es un proceso iterativo en el que a partir de un conjunto inicial de parámetros, $\hat{\theta}_1$, se obtienen los parámetros actualizados, $\hat{\theta}_2$, como el conjunto que maximiza la esperanza de la función verosimilitud sobre la distribución de los factores F_t condicionada a $\hat{\theta}_1$. La distribución de los factores se computa con el suaviazdor de Kalman.

$$\hat{\theta}_2 \;=\; \arg\max_{\theta} E_{F|\hat{\theta}_1}\left[\log L(\theta|X = X_1^T, F)\right]$$

Para motivar el algoritmo EM, Stock y Watson (2002) comienzan mostrando que (3.5) es equivalente a la función log-verosimilitud bajo la condición de que X_{it} sea i.i.d $N(\Lambda_i F_t, 1)$, caso en el que el estimador por mínimos cuadrados es el estimador de máxima verosimilitud gaussiana. Dado que (3.8) es una versión de (3.5) con datos ausentes y que la minimización de (3.5) es computacionalmente simple, el algoritmo EM se puede aplicar para resolver el problema (3.8).

La j-ésima iteración del algoritmo se construye a partir de los estimadores $\hat{\Lambda}$ y $\hat{F}$ construidos en la iteración $j-1$ y de $Q(X, \hat{F}, \hat{\Lambda}, F, \Lambda) = E_{\hat{F},\hat{\Lambda}}[V(F, \Lambda)|X]$, donde $X = \{X_1, \ldots, X_T\}$ denota el conjunto completo de observaciones y $E_{\hat{F},\hat{\Lambda}}[V(F, \Lambda)|X]$ la esperanza de la función log-verosimilitud $V(F, \Lambda)$. Los estimadores de F y Λ en la iteración j resuelven:

$$\min_{F,\Lambda} Q(X, \hat{F}, \hat{\Lambda}, F, \Lambda)$$

Para proceder con los cálculos, nótese que,

$$Q(X, \hat{F}, \hat{\Lambda}, F, \Lambda) = \sum_{i=1}^{N}\sum_{t=1}^{T}\left(E_{\hat{F},\hat{\Lambda}}[X_{it}^2|X] + (\Lambda_i F_t)^2 - 2\hat{X}_{it}(\Lambda_i F_t)\right), \tag{3.9}$$

donde $\hat{X}_{it} = E_{\hat{F},\hat{\Lambda}}[X_{it}|X]$. El primer término en la derecha de (3.9) no depende de F ni Λ, por lo que en aras de minimizar se puede sustituir por $\sum_{i=1}^{N}\sum_{t=1}^{T}\hat{X}_{it}^2$. Esto implica que el F y Λ que minimizan (3.9) se pueden calcular como los términos que minimizan $\hat{V}(F, \Lambda) = \sum_{i=1}^{N}\sum_{t=1}^{T}(\hat{X}_{it} - \Lambda_i F_t)^2$. En la j-ésima iteración, esto se reduce al cálculo usual de CP donde los datos ausentes se sustituyen por su esperanza condicionada a los datos observados y los parámetros de la iteración previa. Si en X se dispone de un subconjunto que constituya un panel balanceado, por ejemplo el panel longitudinal en la Figura 1.1, se puede utilizar para obtener valores iniciales para $\hat{F}$.

Caben considerar algunos detalles para el cálculo de $\hat{X}_{it}$ en varios casos especiales. Sea $X^i = (X_{i1}, \ldots, X_{iT})'$ y x^i el vector de datos observados para X^i. Supongamos que existe una matriz A_i tal que $x^i = A_i X^i$, lo cual se puede hacer en el caso de datos ausentes y agregación temporal por

ejemplo. Entonces $E[X^i|x^i] = F\Lambda_i + A_i'(A_iA_i')^{-1}(x^i - A_iF\Lambda_i)$. Esta formulación es muy general, siendo innecesaria en la mayoría de casos de interés práctico. A partir de ella se llega a aproximaciones más sencillas. Por ejemplo, en el caso donde hay varias observaciones ausentes, en la iteración j los elementos del panel balanceado estimados se pueden construir como $\hat{X}_{it} = X_{it}$ si la observación está disponible y como $\hat{X}_{it} = \hat{\Lambda}_i\hat{F}_t$ en otro caso. El estimador de F se actualiza entonces como el obtenido por CP como en (3.5). El estimador para Λ se calcula con la regresión de $\hat{X}$ sobre la última estimación de F. En caso de tener series con frecuencias temporales distintas, se pueden tratar, por ejemplo, las series trimestrales como series mensuales con datos ausentes en el primer y segundo mes de cada trimestre y proceder como se acaba de exponer.

Una alternativa interesante que nos brinda el algoritmo EM, como indica de Valk *et al.* (2019), es que podemos definir factores para distintos subgrupos de variables, sin tener la limitación de que todos los factores afecten a todas las variables como en los métodos considerados hasta ahora. La única restricción añadida es que el número de shocks de los factores debe ser igual al número de factores. Como ejemplo, supongamos que dividimos un modelo con tres factores en tres grupos (global, nominal y real). El factor global afectará a todas las variables, mientras que los otros dos a las respectivas variables nominales y reales. El MFD considerado tendría la forma:

$$X_t = \begin{pmatrix} \Lambda_{N,G} & \Lambda_{N,N} & 0 \\ \Lambda_{R,G} & 0 & \Lambda_{R,R} \end{pmatrix} \begin{pmatrix} F_t^G \\ F_t^N \\ F_t^R \end{pmatrix} + e_t,$$

donde

$$\begin{pmatrix} \Lambda_{N,G} & \Lambda_{N,N} & 0 \\ \Lambda_{R,G} & 0 & \Lambda_{R,R} \end{pmatrix} = \Lambda$$

$$\begin{pmatrix} F_t^G \\ F_t^N \\ F_t^R \end{pmatrix} = F_t$$

El factor global (F_t^G) se estima considerando todas las variables explicativas, mientras que el nominal (F_t^N) y el real (F_t^R) solo consideran las variables nominales y reales respectivamente. De esta forma pueden estimarse estructuras más flexibles a los datos, así como aumentar la utilidad del MFD imponiendo que ciertos factores actúen sobre un subgrupo determinado. Cuando se contempla

un número elevado de variables es un enfoque interesante, ya que evitamos que ciertas variables puedan ver su efecto camuflado entre las demás. Igualmente, si interesa considerar como variable respuesta una muy relacionada con algún subgrupo de las variables, estimar un factor que contemple solo estas variables será de interés.

Puede resumirse el algoritmo EM como la estimación recursiva de los factores y los valores ausentes repitiendo:

- **Paso E:** Se calcula la esperanza condicionada de la función de verosimilitud a partir de los parámetros estimados en la iteración anterior para el modelo estático, calculada con el suavizador de Kalman
- **Paso M:** Se estiman los nuevos parámetros maximizando la función de verosimilitud calculada en el paso previo.

Se fija algún criterio de convergencia y se repite el algoritmo hasta alcanzarlo.

3.3.2. Estimación en el espacio de estados con datos ausentes

Una forma de solventar los problemas computacionales para el problema de optimización de la función de máxima verosimilitud introducido en la Sección 3.2 es recurrir a la aproximación en dos pasos propuesta por Doz *et al.* (2011), que combina la velocidad de las CP con la eficiencia del filtro de Kalman. En el primer paso, se obtienen estimaciones preliminares de los factores y los parámetros a través del método de CP y se ajusta un modelo a las perturbaciones. Se calculan las cargas por medio de una regresión por mínimos cuadrados y se deduce una estructura para las componentes idiosincráticas a partir de la matriz de covarianzas de los residuos de la regresión. En el segundo paso, los parámetros preliminares se utilizan para construir un modelo de estado-espacio a partir del cual se estima F_t con el filtro de Kalman, estimándolos sobre el panel desbalanceado. Este enfoque tiene también la ventaja de lidiar con paneles dentados y el filtro de Kalman puede suponer un aumento de la eficiencia al "limpiar" en el segundo paso la estimación para los factores, permitiendo una mejor reconstrucción de la dinámica del modelo.

Doz *et al.* (2011) demuestran que para N y T grandes los factores estimados son consistentes para el espacio generado por los factores y robustos ante posibles especificaciones incorrectas de la estructura de las componentes idiosincráticas. Forni *et al.* (2005) muestra la consistencia de las imputaciones que realiza el método en dos pasos para los datos ausentes en el panel. En la sección anterior ya se ha demostrado bajo una serie de condiciones la consistencia de la estimación por CP. De esta forma este método puede interpretarse como no paramétrico al no imponer ninguna suposición previa sobre los parámetros.

Tomemos la notación y escenario de la Figura 1.1. Una propuesta que extiende el algoritmo en dos pasos original, basada en Doz *et al.* (2011) y desarrollada por Cuevas y Quilis (2012) consiste en:

- Estimar los factores estáticos por CP sobre un subconjunto balanceado del panel (donde todas las observaciones estén disponibles).
- Se realiza una regresión de los parámetros excluidos en el paso anterior sobre los factores. Con las nuevas estimaciones se calculan los valores de las observaciones ausentes.
- Se calculan de nuevo los factores con el nuevo panel balanceado para $t = 1, \ldots, T_1$.
- Sobre las estimaciones se aplica el filtro de Kalman en $t = 1, \ldots, T_1$ para los factores comunes, que se proyectan hasta $t = T_2$.
- Los factores obtenidos en el paso anterior se introducen de nuevo en el paso 2 y se itera hasta la convergencia.

El primer panel que se considere debe contar con un número grande de series para ser representativo de las series omitidas en los primeros pasos. Una elección habitual es el subpanel longitudinal en la Figura 1.1.

3.4. Breaks y parámetros variables en el tiempo

Comencemos definiendo qué es un *break*. Se consideran como breaks los cambios de nivel, permanentes o temporales, de la tendencia determinista de una serie temporal. Todas las series son susceptibles de tener varios, y obviar esto en su modelado en caso de que existan puede conducir a errores en la estimación y predicción.

Bajo el MFD, en caso de existir breaks en las series estos serán susceptibles de verse reflejados en la matriz de cargas. Los factores se construyen de forma que capten el mayor porcentaje de variabilidad común a todo el panel, por lo que salvo que todos los breaks de las series se diesen de forma sincrónica (lo cual es improbable en la práctica), por definición los breaks se verán capturados en la componente idiosincrática o como un cambio de nivel en las cargas. Al suponer que las perturbaciones idisincráticas tienen media 0, la única posibilidad que no viola las hipótesis del modelo es la de que las cargas reflejen el cambio de nivel producido por los breaks.

Hasta ahora solo se han considerado MFDs con cargas constantes en el tiempo. Cabe la posibilidad de que en aplicaciones prácticas exista una cierta inestabilidad de los parámetros que debe ser tomada en cuenta. La estimación por CP es robusta ante pequeños breaks, sin embargo si la inestabilidad es amplia se rompe esta propiedad, por tanto es importante comprobar la posible inestabilidad estructural del modelo en las cargas y contar con alternativas en caso de que exista.

3.4.1. Robustez del estimador por Componentes Principales bajo inestabilidad limitada

En este escenario el estimador por CP permanece consistente. Para desarrollar la intuición de esto volvamos al ejemplo de la Sección 3.1 donde consideramos el caso en que solo existe un factor. De nuevo, este argumento se puede generalizar para el caso en que existan más factores. Supongamos que la matriz de cargas de los factores ahora tiene dependencia temporal, por lo que Λ en (2.6) se sustituye por Λ_t. En este caso tendríamos que la media en la sección cruzada de X_t se puede expresar como $\bar{X}_t = \bar{\Lambda}_t F_t + \bar{e}_t$, donde $\bar{\Lambda}_t$ representa la media en la sección cruzada. Sea $\bar{\bar{\Lambda}}$ la media temporal de $\bar{\Lambda}_t$, y por tanto $\bar{X}_t - \bar{\bar{\Lambda}} F_t = (\bar{\Lambda}_t - \bar{\bar{\Lambda}}) F_t + e_t$. Si solo una pequeña fracción de las series tienen un break, o si los breaks en Λ_{it} son incorrelados entre series y tienen dependencia temporal limitada, o si Λ_{it} tiene una dinámica de media cero e incorrelada entre series, por la ley de los grandes números, $\bar{\Lambda}_t - \bar{\bar{\Lambda}} \xrightarrow{\mathbb{P}} 0$ y $e_t \xrightarrow{\mathbb{P}} 0$ y por tanto $\bar{X}_t - \bar{\bar{\Lambda}} F_t \xrightarrow{\mathbb{P}} 0$. De esta forma se concluye que a pesar de la inestabilidad, si $\bar{\bar{\Lambda}} \neq 0$, $\bar{X}_t$ estima de forma consistente el factor salvo un parámetro de escala.

3.4.2. Incorporación de cargas variables en el tiempo y volatilidad

A pesar de la existencia de test para la detección de breaks o de dinámicas en los parámetros, el modelo debe ser ajustado para tener en cuenta tal inestabilidad. Una de las aproximaciones más sencillas es estimar un MFD en cada subconjunto definido por los breaks. El problema de este enfoque es que es demasiado rígido ante dinámicas en los parámetros o breaks en diferentes instantes en varias series.

Una modificación más flexible consiste en modelar la evolución de los parámetros estocásticamente. Si la variación de los parámetros es pequeña, esto puede implementarse en dos pasos. Primero se ajustan los factores por mínimos cuadrados y después se estima un modelo con evolución temporal suponiendo los factores como observados. Los modelos con dependencia temporal han aparecido recientemente en la literatura y no hay muchos trabajos destacados para el caso en que la variación de los parámetros sea elevada. Este debería ser uno de los principales focos de estudio en investigaciones venideras.

Modelando la inestabilidad en las cargas como un proceso estocástico se puede mostrar que si el movimiento no es muy amplio y no hay elevada dependencia entre las series, entonces los resultados de los Teoremas 1 y 2 se mantienen para el escenario con cargas no constantes. El MFD con cargas variables en el tiempo se expresa:

$$\begin{aligned} X_{it} &= \lambda_{it} F_t + e_{it} \\ \lambda_{it} &= \lambda_{i,t-1} + g_i \eta_{it} \end{aligned} \tag{3.10}$$

donde g_i es un escalar y η_{it} es un vector q-dimensional de variables aleatorias. Esta representación implica que la i-ésima carga varía una cantidad $g_i \eta_{it}$ en el instante t. La dependencia limitada entre

series se impone suponiendo que η_{it} tiene una dependencia débil entre las distintas series. Bajo una serie de hipótesis adicionales sobre la variación y dependencia en η_{it}, se puede demostrar que los resultados del Teorema 1 y 2 se siguen cumpliendo en el modelo con factores variables en el tiempo.

3.5. Frecuencias mixtas

Hemos representado todas las series en la frecuencia más alta disponible para la estimación del modelo, pero para realizar predicciones no tiene por qué ser lo más adecuado este enfoque. Por ejemplo, para el forecast del PIB lo habitual es apoyarse en indicadores medidos con una frecuencia mensual. Sea Z_t^Q el valor del PIB en el trimestre t (el superíndice Q indica que se mide de forma trimestral). Interesa tener una predicción que aúne la información para el indicador agregado trimestralmente. Siguiendo Mariano y Murasawa (2003) asumamos que el nivel del PIB como variable trimestral se puede descomponer como la suma de tres observaciones de una variable mensual, Z_t, Z_{t-1} y Z_{t-2}, que indica el crecimiento mes a mes. Así por ejemplo, el nivel del PIB en el tercer trimestre, Z_{III}^Q, es la suma del valor correspondiente a los tres meses del tercer trimestre,

$$Z_{III}^Q = Z_9 + Z_8 + Z_7 \;=\; 3\left(\frac{Z_9 + Z_8 + Z_7}{3}\right),$$

donde Z_{III}^Q representa el dato para el tercer trimestre en la serie del PIB con frecuencia trimestral y Z_i el dato para el mes i de la serie caracterizada con frecuencia mensual.

Mariano y Murasawa (2003) muestran que la media en la derecha de la expresión anterior puede ser aproximada por la media geométrica,

$$Z_{III}^Q = 3\left(Z_9 \cdot Z_8 \cdot Z_7\right)^{1/3} \tag{3.11}$$

Camacho y Pérez-Quiros (2009) indican que las series macroeconómicas son los suficientemente suaves como para que sea adecuada esta aproximación. Emplearemos la notación y desarrollo que siguen Camacho y Doménech (2012). A partir de (3.11) vamos a descomponer los niveles trimestrales como una media ponderada de los niveles mensuales.

Tomando logaritmos en la expresión (3.11):

$$\ln Z_{III}^Q = \ln 3 + \frac{1}{3}(\ln Z_9 + \ln Z_8 + \ln Z_7),$$

lo que nos permite expresar el crecimiento en el tercer trimestre como:

$$\ln Z_{III}^Q - \ln Z_{II}^Q \;=\; \frac{1}{3}(\ln Z_9 + \ln Z_8 + \ln Z_7) \;-\; \frac{1}{3}(\ln Z_6 + \ln Z_5 + \ln Z_4) \;=$$

$$\frac{1}{3}\left[(\ln Z_9 - \ln Z_6) + (\ln Z_8 - \ln Z_5) + (\ln Z_7 - \ln Z_4)\right],$$

y redefiniendo los términos como $Y_t^Q = \ln Z_t^Q - \ln Z_{t-1}^Q$ e $Y_j = \ln Z_j - \ln Z_{j-1}$, tenemos para el crecimiento en el tercer cuarto:

$$Y_{III}^Q = \frac{1}{3}Y_9 + \frac{2}{3}Y_8 + Y_7 + \frac{2}{3}Y_6 + \frac{1}{3}Y_5$$

Y de forma general:

$$Y_t^Q = \frac{1}{3}Y_t + \frac{2}{3}Y_{t-1} + Y_{t-2} + \frac{2}{3}Y_{t-3} + \frac{1}{3}Y_{t-4} \tag{3.12}$$

La expresión (3.12) recibe en la literatura el nombre de *bridge equation* y representa el crecimiento trimestral como una suma ponderada de los ratios de los 5 crecimientos previos. Autores como Camacho y Doménech (2012) implementan en su trabajo este desarrollo demostrando su buen funcionamiento práctico.

Cuevas *et al.* (2017) combinan (3.12) y (2.6) considerada sobre la serie de interés:

$$Y_t \;=\; \Lambda_Y F_{Y,t} + e_{Y,t}$$

donde el subíndice Y indica que consideramos los parámetros relativos a la serie Y_t. Con estas dos ecuaciones expresan la serie trimestral a partir de la mensual como:

$$Y_t^Q = \frac{1}{3}\Lambda_Y F_{Y,t} + \frac{2}{3}\Lambda_Y F_{Y,t-1} + \Lambda_Y F_{Y,t-2} + \frac{2}{3}\Lambda_Y F_{Y,t-3} + \frac{1}{3}\Lambda_Y F_{Y,t-4} +$$
$$\frac{1}{3}e_{Y,t} + \frac{2}{3}e_{Y,t-1} + e_{Y,t-2} + \frac{2}{3}e_{Y,t-3} + \frac{1}{3}e_{Y,t-4}$$

Con un enfoque similar, pero en este caso sobre los factores ya estimados y no la propia variable, Cuevas y Quilis (2012), siguiendo la línea sugerida en Mariano y Murasawa (2003), convierten los factores estimados con frecuencia mensual a "factores trimestrales" según:

$$f_t^Q = \left(\frac{1}{3} + \frac{2}{3}L + L^2 + \frac{2}{3}L^3 + \frac{1}{3}L^4\right) f_t \tag{3.13}$$

Capítulo 4

Número de factores y dimensión del panel

La selección del número de factores se puede hacer en base a un conocimiento previo, en base a una herramienta visual o utilizando un criterio de información. Notemos que hasta ahora se ha desarrollado la mayor parte del trabajo suponiendo que el número de factores estáticos (q) o dinámicos (r) son conocidos, por tanto será necesario contar con un estimador consistente de este parámetro a fin de que las sucesivas estimaciones no conduzcan a error debido a considerar un número erróneo de factores latentes. Ha quedado claro la importancia de este parámetro en cuanto a que todas las estimaciones y cálculos se apoyan en él, de ahí que sea un punto clave su correcta estimación.

4.1. Estimación del número de factores estáticos

El más utilizado y destacado criterio en la literatura es el propuesto en Bai y Ng (2002). Hay más alternativas de las aquí recogidas, como la aproximación sugerida por Onatski (2009, 2010) que se centra en la diferencia entre los autovalores k y $k+1$ de Σ_X o el ratio entre estos y selecciona el que lo maximice. Aproximaciones como esta son muy recientes y no hay estudios prácticos que estudien su funcionamiento práctico, por lo que no se profundizará en estas técnicas. Diferentes métodos tienden a producir distintos resultados, por lo que parece un importante campo de desarrollo sobre el que hay un escueto número de alternativas. En el marco donde se encuentra este trabajo, los criterios que a continuación se recogen han demostrado tener un buen desempeño y ser consistentes, por lo que nos centraremos en ellos para su posterior aplicación.

Los primeros criterios se basan en hipótesis con N y T fijos, lo cual es poco justificable conforme

han ido creciendo el número de series disponibles los últimos años. Stock y Watson (1998) proponen una modificación del BIC, con la restricción de que es necesario que $N >> T$ y que solo es apropiado con el objetivo de realizar predicciones, al centrarse en el número de factores para la predicción y no en modelar la dinámica de X_t. El criterio de Bai y Ng (2002) solventa los problemas de los criterios previos aportando uno en línea con los clásicos y que ha demostrado un buen funcionamiento. Podría decirse que esta propuesta es la generalización al MFD de los criterios de información clásicos, como los empleados por ejemplo en la regresión lineal para la selección del número de parámetros.

4.1.1. Gráfico de sedimentación

Suele referirse a este gráfico por su nombre en inglés, *scree plot*. Es una representación de la contribución de la k-ésima componente principal a la media del coeficiente R^2 de N regresiones de X_t sobre las k primeras CP estimadas del panel. En el caso donde no hay datos ausentes, el scree plot consiste en una gráfica de los autovalores ordenados de Σ_X normalizados entre la suma de todos los autovalores. Este procedimiento es análogo al que se utiliza en CP usuales, donde se busca un "codo" en el gráfico para seleccionar el número de CP a emplear en el análisis. Se busca seleccionar el número mínimo de factores que expliquen un gran porcentaje de los movimientos de X_t y que cualquier factor añadido tenga poco poder explicativo extra. Este enfoque tiene el problema de estar sujeto a cierto grado de subjetividad y la incomodidad de no poder ser automatizado.

4.1.2. Criterio de información

En la literatura se desarrollan este tipo de criterios al interesar una medida objetiva y no una representación sujeta a interpretaciones subjetivas como el scree plot, el cual en algunos casos no conduce a una elección clara. Los criterios de información para el MFD se basan en el clásico problema entre ajuste y varianza. Uno de los más comunes en estadística es el AIC, que introduce una penalización sobre el número de parámetros para buscar una compensación entre el beneficio explicativo de añadir un parámetro más respecto al aumento de variabilidad que esto conlleva. Bai y Ng (2002) partiendo del MFD estático extienden este criterio incluyendo un término que penaliza la suma de errores al cuadrado y el número de parámetros:

$$IC(k) = \ln V_k(k, \hat{F}^k) + kg(N,T), \tag{4.1}$$

donde $\hat{F}^k$ denota el vector de factores estáticos estimado con dimensión k, $V_k(k, \hat{F}^k)$ es la función objetivo por mínimos cuadrados (3.5), suponiendo que el número de factores estáticos estimados es k y $g(N,T)$ una función de penalización tal que $g(N,T) \to 0$ y $\text{mín}(N,T)g(N,T) \to \infty$ conforme $N,T \to \infty$. Se considera $k < \text{mín}\{N,T\}$. El escalar que minimice la función (4.1) será el que se tome como estimador del número de factores a considerar en el modelo.

Además de dotarnos de un criterio objetivo e intuitivo, otra gran innovación del criterio propuesto por Bai y Ng (2002) es que no se imponen restricciones sobre N o T, a diferencia de los criterios previos. Los resultados también se sostienen bajo heterocedasticidad temporal y/o en las series e incluso bajo correlación débil entre las variables. Se permite además dependencia débil entre los factores y las perturbaciones idiosincráticas:

$$E\left[\frac{1}{N}\sum_{i=1}^{N}\left\|\frac{1}{\sqrt{T}}\sum_{t=1}^{T}F_t e_{it}\right\|^2\right] \leq M, \quad M \ \textit{constante t.q.} \ M < \infty \tag{4.2}$$

Como sucedía antes, el elevado número de series que se consideran hace que el modelo exacto sea susceptible de no cumplirse. Bai y Ng (2002) asumen que en caso de existir correlación entre los errores, está limitada. En caso de independencia de los factores y las componentes idosincráticas, una hipótesis estándar en el MFD, la condición (4.2) se cumple implícitamente por FC2, FC4 y M.

Supongamos que los factores son observables pero no las cargas. El problema entonces se reduciría a escoger los k factores que capturasen mejor los movimientos de X_t y estimar las cargas correspondientes. Como el modelo es lineal y los factores conocidos, Λ_i se podría estimar por mínimos cuadrados para $i = 1, \ldots, N$. Estamos entonces ante un clásico problema de selección de variables y cobra sentido un criterio de selección análogo adaptado al MFD, como (4.1).

Una vez motivada la idea, Bai y Ng (2002) buscan una función de penalización $g(N,T)$ tal que un criterio de la forma

$$PC(k) = V(k, \hat{F}^k) + kg(N,T), \tag{4.3}$$

pueda estimar consistentemente q.

Teorema 3. *(Teorema 2 de Bai y Ng (2002)). Sea kmax un entero tal que $q \leq kmax$. Supongamos que se cumplen las hipótesis FC, M y (4.2) y que se estiman k factores por CP. Sea $\hat{k} = \arg min_{0\leq k\leq kmax} PC(k)$. Entonces,*

$$\lim_{N,T\to\infty} \mathbb{P}[\hat{k} = q] = 1 \quad si:$$

- $g(N,T) \to 0$
- $C_{NT}^2 g(N,T) \to \infty$ *conforme* $N,T \to \infty$, *donde* $C_{N,T} = \min\left\{\sqrt{N}, \sqrt{T}\right\}$.

Estas dos últimas condiciones son necesarias en el sentido de que si alguna no se cumple, puede existir un modelo que satisfaga las condiciones FC, M y (4.2) y que el número de factores estimado no sea consistente. De todas formas no siempre se requieren. Como Corolario al Teorema 3, bajo las mismas hipótesis, (4.1) también estima de forma consistente q.

Corolario 1. *(Corolario 1 en Bai y Ng (2002)). Bajo las hipótesis del Teorema 3, un criterio de la forma:*

$$IC(k) = \ln(V(k, \hat{F}^k)) + kg(N, T)$$

estima también de forma consistente q.

Notemos que (4.1) y (4.3) son una clara generalización de los criterios de información clásicos. El término $V(k, \hat{F})$ es la varianza residual media cuando se estiman k factores estáticos y el segundo penaliza la inclusión de más factores en el modelo.

Llegados a este punto, la única cuestión que queda por resolver es la selección de una función $g(N, T)$ que cumpla las condiciones del Teorema 3. Bai y Ng (2002) proponen 3 funciones en su trabajo que han sido aplicadas en la mayoría de trabajos posteriores, demostrando que la consistencia teórica se traslada a la aplicación práctica satisfactoriamente. Sea $\hat{\sigma}^2$ un estimador consistente de $(NT)^{-1} \sum_{i=1}^{N} \sum_{t=1}^{T} E[e_{it}]^2$. Los 3 criterios propuestos por Bai y Ng (2002) son:

$$PC_{p1}(k) = V(k, \hat{F}^k) + k\hat{\sigma}^2 \left(\frac{N+T}{NT} \right) \ln \left(\frac{NT}{N+T} \right); \tag{4.4}$$

$$PC_{p2}(k) = V(k, \hat{F}^k) + k\hat{\sigma}^2 \left(\frac{N+T}{NT} \right) \ln C_{NT}^2; \tag{4.5}$$

$$PC_{p3}(k) = V(k, \hat{F}^k) + k\hat{\sigma}^2 \left(\frac{\ln C_{NT}^2}{C_{NT}^2} \right) \tag{4.6}$$

El Corolario 1, lleva a considerar también los siguientes criterios:

$$IC_{p1}(k) = \ln \left(V(k, \hat{F}^k) \right) + k \left(\frac{N+T}{NT} \right) \ln \left(\frac{NT}{N+T} \right); \tag{4.7}$$

$$IC_{p2}(k) = \ln \left(V(k, \hat{F}^k) \right) + k \left(\frac{N+T}{NT} \right) \ln C_{NT}^2; \tag{4.8}$$

$$IC_{p3}(k) = \ln \left(V(k, \hat{F}^k) \right) + k \left(\frac{\ln C_{NT}^2}{C_{NT}^2} \right) \tag{4.9}$$

La ventaja de estos criterios es que no dependen de la elección de $kmax$ ni de $\hat{\sigma}^2$, lo cual es de interés en la práctica. Stock y Watson (2005) comprueban que estos criterios estiman correctamente el número de factores a partir de técnicas de Monte Carlo, donde como mínimo aciertan un 88 % de las veces y con un error de 1 o menos el 99 %.

Bai y Ng (2002) explican que por construcción los criterios permiten que F_t tenga algún tipo de dependencia de forma que $\Phi(L)F_t = \epsilon_t$, siendo $\Phi(L)$ una matriz de polinomios de retardo. Sin embargo, no se considera el caso en que las dinámicas entran en X_t directamente. Si alguno de los criterios se aplican a algún modelo con esa estructura, el estimador para el número de factores actúa como cota superior para el verdadero. Si consideramos un modelo donde $X_{it} = a_i f_t + b_i f_{t-1} + e_{it}$, desde el punto

de vista dinámico solo existe un factor, pero los criterios anteriores lo tratarían como si existiesen el doble de los que realmente hay, salvo que el rango de la matriz de cargas sea de rango uno. De ahí que en próximas secciones consideremos shocks en los factores y su correcta estimación, para poder diagnosticar estos escenarios y ser consecuentes en la estimación del número de factores.

4.2. Estimación del número de factores dinámicos

El número de factores dinámicos puede ser menor que el de factores estáticos, en cuanto $r \leq q$. En caso de cumplirse la desigualdad estrictamente los factores estáticos siguen un proceso singular, en cuanto a que la matriz de covarianzas de las innovaciones de F_t ($G\eta_t$ en (2.7)) es singular de rango $r < q$. Esto implica que la matriz espectral de F_t es singular. Para lidiar con esta situación es posible estimar r dado q, estimando el rango de la matriz de covarianzas de los errores del modelo VAR ajustado sobre los factores estáticos. Otra alternativa es recurrir a un criterio de información basado en la verosimilitud del MFD dinámico, pero es preferible la primera aproximación para evitar el cálculo de dicha verosimilitud.

En la literatura se emplea el término *número de shocks* para referirse al número de factores dinámicos, ya que en terminología de series de tiempo se dice que por como están construidos, los factores estáticos están conducidos por los shocks de los factores dinámicos.

Bai y Ng (2007) estiman el rango de la matriz de covarianzas del modelo VAR ajustado a los factores estáticos utilizando las q primeras CP estimadas a partir de uno de los criterios de información introducidos en la Sección 4.1.2. Sea $\hat{F}_t$ el vector con los q factores estáticos estimado por CP y $\hat{\mu}_t = \hat{G}\hat{\eta}_t$ los residuos del VAR (2.7). Introducimos la notación con μ_t para simplificar las expresiones. La idea es comprobar si los autovalores de la matriz de covarianza de $\hat{\mu}_t$, $\hat{\Sigma}_\mu$, son distintos de 0, lo que indica el rango de $\hat{\Sigma}_\mu$ y por consiguiente el número de shocks que conducen a los factores estáticos, i.e. el número de factores dinámicos subyacentes. Numéricamente, se testa si un autovalor dado está por debajo de un nivel de tolerancia. Para ello, definamos los autovalores $c_1 \geq c_2 \geq \cdots \geq c_q \geq 0$ de $\hat{\Sigma}_\mu$ y la k-ésima normalización del k-ésimo autovalor:

$$\hat{D}_k = \left(\frac{c_k^2}{\sum_{j=1}^{q} c_j^2} \right)^{1/2}$$

Para algún $0 < m < \infty$ y $0 < \delta < 1/2$ que fije el nivel de tolerancia, se define el vector K:

$$K = \left\{ k : \hat{D}_k < m \,\text{mín}[N^{\frac{1}{2}-\delta}, T^{\frac{1}{2}-\delta}] \right\}$$

Bai y Ng (2007) indican que $m = 1$ y $\delta = 0.1$ conducen a buenas estimaciones del número de factores dinámicos. El número de shocks de los factores se toma $\hat{r} = \text{mín}\,\{k \in K\}$. Este estimador converge en probabilidad al número real de shocks si q es el verdadero número de factores estáticos, i.e.

$\hat{r} \xrightarrow{\mathbb{P}} r$ según $T \to \infty$. En la sección anterior ya se comentó el buen funcionamiento de los estimadores propuestos por Bai y Ng (2002) para determinar q y por tanto $\hat{r}$ puede precisarse en la práctica.

El estimador de la matriz de covarianzas de μ_t se toma $\hat{\Sigma}_\mu = \frac{1}{T}\sum_{t=1}^{T} \hat{\mu_t}'\hat{\mu_t}$. Bai y Ng (2007) demuestran en su trabajo que, siendo $H \in \mathcal{M}_{q\times q}$ de rango completo una matriz de rotación de forma que $\hat{F}_t$ estime HF_t, se tiene:

$$\hat{\Sigma}_\mu - H\Sigma_\mu H' = O_p\left(1/\min\left\{\sqrt{N}, \sqrt{T}\right\}\right)$$

La estimación del número de factores dinámicos anterior sería válida si F_t fuese observable, ya que con la descomposición espectral de la matriz de covarianzas de los errores de su representación como modelo VAR podríamos determinar r. Lo que impide el análisis anterior es que ni F_t ni el número de factores estáticos q son observados. Bai y Ng (2007) a partir de un razonamiento similar al anterior proponen dos nuevos estimadores teniendo en cuenta esto.

Se considera el número de factores estáticos q dado y se estima un modelo VAR de orden P sobre los factores, donde el orden P se selecciona por el criterio BIC. Esto no supone una pérdida de generalidad ya que como vimos anteriormente es posible la estimación del número de factores estáticos de forma robusta. A continuación se calcula la descomposición espectral de la matriz de covarianzas de los residuos estimados del modelo VAR, $\hat{\Sigma}_\mu$, de dimensión $(q \times q)$. Para $l = 1, \cdots, q-1$ Bai y Ng (2007) introducen:

$$\hat{D}_{1,l} = \left(\frac{\hat{c}_{l+1}}{\sum_{j=l+1}^{q}\hat{c}_j}\right)^{1/2}$$

$$\hat{D}_{2,l} = \left(\frac{\sum_{j=l+1}^{q}\hat{c}_j}{\sum_{j=1}^{q}\hat{c}_j}\right)^{1/2},$$

donde $\hat{D}_{1,l}$ representa una medida de la contribución marginal del $(l+1)$-ésimo autovalor y $\hat{D}_{2,l}$ de la contribución acumulada de los autovalores, bajo la hipótesis nula de que $rango(\hat{\Sigma}_\mu) = r$ y que $c_l = 0$ para $l > r$.

Así, de acuerdo a la contribución marginal que consideremos, la estimación del número de factores dinámicos $\hat{r}$ se obtiene tomando el mínimo de:

$$\left\{l \mid \hat{D}_{1,l} \leq \frac{c}{\min\left\{n^{2/5}, T^{2/5}\right\}}\right\}$$

o

$$\left\{l \mid \hat{D}_{2,l} \leq \frac{c}{\min\left\{n^{2/5}, T^{2/5}\right\}}\right\}$$

Bai y Ng (2007) sugieren emplear $c = 1$ en base a simulaciones de Monte Carlo.

Aplicación práctica Barhoumi *et al.* (2018) recopilan todos estos criterios e indican cómo se procede ante un caso práctico:

1. Emplear el criterio de Bai y Ng (2002) para estimar el número de factores estáticos q.
2. Ajustar un modelo VAR(P) a los q factores estimados, seleccionando el orden P minimizando el criterio BIC.
3. Con el criterio de Bai y Ng (2007) estimar la matriz de covarianzas de los residuos μ_t del modelo VAR(P) para obtener el número óptimo de factores dinámicos r.

4.3. Número de series

Hasta ahora se ha supuesto que las N variables introducidas en el modelo tienen valor informativo para la estimación de los factores. La motivación principal del MFD es el poder utilizar un gran número de series como conjunto de datos. Uno de los dilemas de esto es que al introducir muchas variables el panel es susceptible de incorporar alguna con un comportamiento extraño o extremo que puede ser mejor eliminar para evitar que afecte en las estimaciones. Una posibilidad es utilizar un método de pre-cribado para reducir el conjunto de datos, como se sugiere en Bai y Ng (2008). Se indica que estudiar la importancia de la componente común en cada serie puede ayudar a eliminar variables poco informativas o redundantes para una subsiguiente regresión más eficiente. Otra alternativa es eliminar variables muy dispersas, volátiles o infrecuentes en el sentido de tener un elevado porcentaje de observaciones ausentes.

Un modelo VAR habitualmente modela entre 6 y 10 series para que sea viable computacionalmente. El problema de esto, como se ha comentado, es el posible sesgo que introducimos al dejar un número elevado de series sin considerar en el modelo. Esto concuerda con el clásico balance que se debe hacer entre sesgo y varianza a la hora de realizar cualquier ajuste. Hasta ahora hemos considerado modelos donde N se supone grande, pero, ¿realmente es necesario considerar un extenso número de series para mejorar la precisión? Y aún más importante, ¿estamos incrementando la varianza demasiado al emplear este enfoque? Stock y Watson (2002a) mostraron que la mejora del ajuste (en términos de MSE) cuando N aumenta a partir de 50 es despreciable en datos simulados y sobre el forecast para el índice de producción industrial estadounidense, realizado a partir de 149 variables medidas entre 1959 y 1970. Bai y Ng (2002) a partir de simulaciones muestran que se puede estimar el número de factores latentes con precisión tomando N del orden de 40 si los errores son independientes. Todo esto nos sugiere que N no tiene porqué ser extremadamente grande para obtener estimaciones precisas de los parámetros del MFD por el método de CP.

La teoría asintótica introducida hasta ahora supone una baja correlación cruzada entre los errores. Conforme aumentamos el número de series, crece la probabilidad de contar con errores correlados.

Además, una característica importante de las componentes comunes es que capten la dinámica del grueso de las series. Conforme se introducen más aumenta la probabilidad de introducir series con demasiado ruido, lo que hace que los pesos de los factores se vayan reduciendo y conduce a una aproximación más pobre y menos interesante.

Como ya se ha visto, podemos descomponer la matriz de covarianzas de los datos como $\Sigma_X = \Sigma_{\Lambda F} + \Sigma_e$. Como F_t es común a todas las variables, $\Sigma_{\Lambda F}$ tiene q autovalores no nulos que incrementan conforme aumenta N. Una característica fundamental del MFD es que los q mayores autovalores de Σ_X también aumentan con N. Esto es lo que conduce a suponer que el espacio generado por los factores puede ser estimado a partir de la descomposición espectral de la matriz de covarianzas muestral de X_t.

Para desarrollar la intuición en por qué el aumentar el número de variables puede ser contraproducente consideremos un ejemplo (extraído de Boivin y Ng (2006)) donde se considera un factor estático ($q = 1$) y que las cargas son idénticas ($\Lambda_i = \lambda \ \forall i$). Para un cierto N, tenemos que $\hat{F}_{t,N} = F_t + \frac{1}{N}\sum_{i=1}^{N} e_{it}$, de donde se sigue que $\text{var}(\hat{F}_{t,N}) = \text{var}(\frac{1}{N}\sum_{i=1}^{N} e_{it})$. Si e_{it} son iid, $\text{var}(\hat{F}_{t,N}) = \sigma^2/N$, que decrecería conforme aumenta N. Este resultado es análogo al de la regresión clásica cuando las cargas son observadas. Sin embargo, al igual que en el caso de la regresión clásica, los estimadores por mínimos cuadrados no tienen una relación inequívoca entre error y N cuando se relaja la hipótesis de errores iid.

Consideremos la estimación de la media muestral. Supongamos N_1 series recogidas de una población con varianza σ_1^2, a partir de la que calculamos la media de la variable a predecir, $\bar{y}$. Supongamos un conjunto de N_2 series de una población con varianza σ_2^2, con $\sigma_1^2 < \sigma_2^2$. Con las $N = N_1 + N_2$ series calculamos la media muestral de nuevo, $\tilde{y}$. Se tiene que $\text{var}(\tilde{y})/\text{var}(\bar{y}) > 1$ si $(N_1\sigma_1^2 + N_2\sigma_2^2)/N^2 > \sigma_1^2/N_1$. Entonces el aumento en la eficiencia de la estimación depende también de las propiedades de las series que añadimos. Consideremos el ejemplo donde por error se introducen las N_1 series dos veces. De esta forma $N = 2N_1$ y N_1 pares de errores idiosincráticos están perfectamente correlados. Si los N_1 errores son iid, $\text{var}(\hat{F}_{t,N}) = \frac{\sigma^2}{N_1}$, que depende de N_1 y no de N. No se ha ganado nada al añadir las series ya que las series duplicadas aumentan la variación de la componente común pero no se reduce la de los errores.

Boivin y Ng (2006) concluyen que debido al desarrollado de la teoría enfocado para N grande, hay una tendencia a considerar tantas series como sea posible, lo cual no siempre puede ser el mejor enfoque como acabamos de ver. En simulaciones hechas por ellos mismos y en trabajos prácticos comprueban que un N del orden de 40 parece no producir peores estimaciones e incluso en muchos casos mejoran a las realizadas con un número mayor de series. Por el momento no existe ninguna guía o consenso indicando qué variables introducir en el modelo. Lo que se recomienda en Boivin y Ng (2006) es no centrarse solo en el número de datos en sí si no también en la calidad de estos, buscando un conjunto de series poco correladas y con dinámicas importantes de cara a la estimación de los factores.

Capítulo 5

Inferencia sobre el modelo

Hasta el momento hemos supuesto que se cumplían unas serie de condiciones para la consistencia del modelo y poder considerar como despreciable el error en la estimación de los factores. Si este error no se puede suponer despreciable, se necesitan las propiedades de la distribución de los estimadores. Además, será de utilidad construir intervalos de confianza para los estimadores. En este capítulo se sigue en la línea donde N y T se suponen grandes y se amplía el escenario permitiendo N grande y que T sea fijo, caso donde se necesitará ortogonalidad y homocedasticidad asintótica (esto se precisa a continuación). En el escenario con N y T grandes se puede establecer la consistencia en presencia de cierta correlación serial y heterocedasticidad. Bai (2003) recalcan que las hipótesis clásicas del MFD como N fijo y menor que T, la independencia de los errores e_{it} en la sección temporal y cruzada o la independencia de F_t y e_t son muy restrictivas en el marco económico. Como ya se comentó, estas limitaciones suponen la introducción del MFD aproximado y más adelante modelos donde se permite cierta correlación entre los errores idiosincráticos e incluso heterocedasticidad.

A continuación se recogen varios resultados y teoremas recogidos en Bai (2003) supuestas una serie de hipótesis. La principal motivación de introducir estos resultados es que dotan al investigador con tasas de convergencia y distribuciones asintóticas, las cuales se pueden emplear para determinar intervalos de confianza por ejemplo. Estos hacen referencia a la estimación del MFD por CP, al ser como se comentó la técnica más empleada en la literatura. No se profundiza en la demostración ya que el objetivo de esta sección es mostrar las propiedades de los estimadores introducidos para tener una noción y poder emplear los resultados más adelante.

5.1. Propiedades de los estimadores

Las siguientes suposiciones amplían las introducidas anteriormente y algunas simplemente extienden hipótesis previas. Se introducen conforme son necesarias para mostrar los revestimientos que se van imponiendo para llegar a los distintos resultados.

Suposiciones A Se aplica en Bai y Ng (2002) (más la hipótesis FC4) para estimar el número de factores de forma consistente,

$$E||F_t^0||^4 < \infty,$$

donde F_t^0 representa los verdaderos factores.

Suposiciones C Amplían las hipótesis M. Limitan la dependencia temporal y en la sección cruzada, así como las heterocedasticidad.

Existe una constante positiva $M < \infty$ t.q. $\forall N, T$,

1. $E(e_{it}) = 0, \quad E|e_{it}|^8 \leq M$
2. $E(e_{it}e_{jt}) = \tau_{ij,ts}$ y $(NT)^{-1} \sum_{i=1}^{N} \sum_{t=1}^{T} |\tau_{ij,ts}| \leq M$

El primer resultado presenta la consistencia uniforme de los factores estimados por CP.

Proposición 1. *(Proposición 2 en Bai (2003)). Bajo las hipótesis A, FC4, FC3, FC1, C, M, 4.2*

$$\max_{1 \leq T} \left|\left| \hat{F}_t - H' F_t^0 \right|\right| \; = \; O_p(T^{1/2}) + O_p((T/N)^{1/2}),$$

donde $\hat{F}_t$ el estimador por CP y H es una matriz de rango completo. La proposición se da en función de una matriz H arbitraria en cuanto a que los factores no son separablemente identificables y por tanto solo podemos esperar estimar correctamente el espacio generado por ellos. H es tal que $\hat{F}_t$ es el estimador de $F_t^0 H$. En la aplicación práctica al interesar explicar las series en función de los factores no supone ninguna restricción, ya que conocer el espacio generado es suficiente para el ajuste de las series respecto a los factores.

La Proposición 1 aporta una cota superior para la máxima desviación de los factores estimados respecto a los verdaderos. Bai (2003) indica que puede precisarse la cota incluso más en caso de que $\liminf N/T^2 \geq 0$, en donde la máxima desviación es $O_p(T^{-1/2})$, lo cual es un resultado a destacar.

Suposiciones F Sobre los momentos y empleando el Teorema Central del límite.

Existe un $M < \infty$ t.q. $\forall N, T$,

1. Para cada $t \in 1, \cdots T$, $\quad E||\frac{1}{\sqrt{NT}} \sum_{s=1}^{T} \sum_{k=1}^{N} F_s^0 (e_{ks} e_{kt} - E[e_{ks} e_{kt}])||^2 \leq M$
2. $E||\frac{1}{\sqrt{NT}} \sum_{t=1}^{T} \sum_{k=1}^{N} F_t^0 \lambda_k^{0'} e_{kt}||^2 \leq M$
3. Para cada t, conforme $N \to \infty$, $\quad \frac{1}{\sqrt{NT}} \sum_{i=1}^{N} \lambda_i^0 e_{it} \xrightarrow{d} N(0, \Gamma_t)$,

   donde $\Gamma_t = \lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{N} \lambda_i^0 \lambda_j^{0'} E[e_{ij} e_{jt}]$
4. Para cada i, según $T \to \infty$, $\frac{1}{\sqrt{T}} \sum_{t=1}^{T} F_t^0 e_{it} \xrightarrow{d} N(0, \Upsilon_i)$,

   donde $\Upsilon_i = \text{plím}_{T\to\infty} \frac{1}{T} \sum_{s=1}^{T} \sum_{t=1}^{T} E[F_t^0 F_s^{0'} e_{is} e_{it}]$

Otro resultado destacado es la distribución límite de las componentes comunes estimadas. Sea $C_{it}^0 = F_t^{0'} \lambda_i^0$ la componente común verdadera para la variable i-ésima y $\hat{C}_i t = \hat{F}_t' \hat{\lambda}_i$ su estimador por CP.

Teorema 4. *(Teorema 3 de Bai (2003)). Bajo las hipótesis A, FC4, FC3, FC1, C, M, 4.2 y F, conforme $N, T \to \infty$, para cada i y t,*

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

donde $V_{it} = \lambda_i^{0'} \Sigma_\Lambda^{-1} \Gamma_t \Sigma_\Lambda^{-1} \lambda_i^0$ y $W_{it} = F_t^{0'} \Sigma_F^{-1} \Upsilon_i \Sigma_F^{-1} F_t^0$.

Por tanto las componentes comunes estimadas son siempre asintóticamente normales, con ratio de convergencia $\delta_{NT} = \text{mín}\left\{\sqrt{N}, \sqrt{T}\right\}$. Para verlo, basta expresar el Teorema 4 como:

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

El denominador está acotado inferior y superiormente, así el ratio de convergencia es δ_{NT}. Cabe recordar que el ratio de convergencia da una idea de la "velocidad" a la que una serie convergente se aproxima a su límite. Esto deja patente la importancia que cobra el tamaño de N y T en estos resultados.

En Bai (2003) se demuestra que el ratio de convergencia de $\hat{F}_t$ es $\text{mín}\left\{\sqrt{N}, T\right\}$. Si T permanece fijo esto implica que $\hat{F}_t$ no es consistente, por lo que se deben imponer más condiciones para tener la consistencia del estimador para T fijo.

Teorema 5. *(Teorema 4 de Bai (2003)). Asumamos las condiciones A, FC4, FC3, FC1, C, M, 4.2, F. Bajo T fijo, una condición necesaria y suficiente para la consistencia es la ortogonalidad asintótica y la homocedasticidad asintótica*

El Teorema anterior indica que para T fijo no es posible una estimación consistente en presencia de correlación serial y heterocedasticidad.

Suposiciones H Esta hipótesis implica la homocedasticidad y ausencia de correlación serial.

$E[e_{it}e_{is}] = 0$ si $t \neq s$, $E[e_{it}^2] = \sigma_i^2$ y $E[e_{it}e_{jt}] = \tau_{ij}$, $\forall t, i, j$.

Sea V_{NT} la matriz diagonal compuesta por los q mayores autovalores de la matriz $\frac{1}{NT}XX'$. En Bai (2003) se demuestra que $V_{NT} \to V$, una matriz definida positiva. Definamos $D_{NT} = V_{NT}(V_{NT} - \frac{1}{T}\bar{\sigma^2}_N)^{-1}$, con $\bar{\sigma}_N^2 = \frac{1}{N}\sum_{i=1}^N \sigma_i^2$; entonces $D_{NT} \to I_q$ según $N, T \to \infty$. Sea $\bar{H} = HD_{NT}$ y Q la matriz $(q \times q)$ invertible tal que $plim_{T,N\to\infty} \frac{\hat{F}'F^0}{T} = Q$.

Teorema 6. *(Teorema 5 de Bai (2003)). Bajo las condiciones A, FC4, FC3, FC1, C, M, 4.2, F, H, según $N, T \to \infty$,*

$$\sqrt{N}(\hat{F}_t - \bar{H}'F_t^0) \xrightarrow{d} N(0, V^{-1}Q\Gamma Q'V^{-1}),$$

donde $\Gamma = plím(\Lambda^{0'}\Sigma_e\Lambda^0/N)$, $\Sigma_e = E(e_te_t') = (\tau_{ij})$ y $\bar{H}$ es la matriz escalada a partir de la matriz de covarianza de los datos y tal que $\bar{H} \xrightarrow{\mathbb{P}} H$.

5.2. Implementación práctica

Los teoremas y resultados anteriores nos sitúan en un marco donde podemos realizar inferencia sobre la estimación de los factores y las componentes comunes. Están dados en función de las matrices de covarianza de las verdaderas variables, inobservables en la práctica, por lo que son necesarios estimadores consistentes.

Matriz de covarianza de los factores estimados La matriz de covarianza de $\hat{F}_t$ está dada por $\Pi_t = V^{-1}Q\Gamma Q'V^{-1}$, es decir,

$$\Pi_t = \text{plím } V_{NT}^{-1}\left(\frac{\hat{F}'F^0}{T}\right)\left(\frac{1}{N}\sum_{i=1}^N \sigma_{it}^2\lambda_i^0\lambda_i^{0'}\right)\left(\frac{F^{0'}\hat{F}}{T}\right)V_{NT}^{-1}$$

La matriz anterior contiene a F^0 y Λ^0, que pueden ser reemplazados por su correspondiente esti-

mador $\hat{F}$, $\hat{\Lambda}$. Un estimador consistente de la matriz de covarianza viene dado por:

$$\hat{\Pi}_t = V_{NT}^{-1} \left(\frac{\hat{F}'\hat{F}}{T} \right) \left(\frac{1}{N} \sum_{i=1}^{N} \hat{e}_{it}^2 \hat{\lambda}_i \hat{\lambda}_i' \right) \left(\frac{\hat{F}'\hat{F}}{T} \right) V_{NT}^{-1} \tag{5.1}$$

donde $\hat{e}_{it} = X_{it} - \hat{\lambda}_i \hat{F}_t$.

Matriz de covarianza de las cargas estimadas Sea $\Omega_i = \text{plím}_{T\to\infty} \frac{1}{T} \sum_{s=1}^{T} \sum_{t=1}^{T} E[F_t^0 F_s^{0'} e_{is} e_{it}]$. La matriz de covarianzas asintótica de $\hat{\lambda}_i$ está dada por:

$$\Theta_i = (Q')^{-1} \Omega_i Q^{-1}$$

Matriz de covarianza de las componentes comunes estimadas Los estimadores para V_{it} y W_{it} con un razonamiento análogo al anterior son:

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

$$\hat{W}_{it} = \hat{F}_t' \left(\frac{\hat{F}'\hat{F}}{T} \right)^{-1} \hat{\Theta}_i \left(\frac{\hat{F}'\hat{F}}{T} \right)^{-1} \hat{F}_t \equiv \hat{F}_t' \hat{\Theta}_i \hat{F}_t$$

Teorema 7. *(Teorema 6 de Bai (2003)). Supongamos las hipótesis A, FC4, FC3, FC1, C, M, 4.2, F e independencia en la sección cruzada. Entonces según $T, N \to \infty$, $\hat{\Pi}$, $\hat{\Theta}_i$, $\hat{V}_{it}$ y $\hat{W}_{it}$ son estimadores consistentes de las respectivas variables.*

Capítulo 6

Aplicaciones de los factores estimados

Una vez estimados los factores por medio de alguna de las técnicas introducidas, pueden ser empleados para sucesivas regresiones sobre variables de interés o incluso para enriquecer un modelo VAR introduciéndolos como variables. Los modelos VAR son uno de los más empleados en econometría, pero tienen los problemas ya comentados de restringir el número de series que pueden modelar. Dado que los factores condensan los movimientos principales de un elevado número de variables, resulta natural considerar sustituir el grueso de las variables por los factores y retener solo las variables de interés. En la parte práctica de este trabajo la principal motivación para la estimación de los factores es utilizarlos como variables independientes que nos proporcionen una representación de una variable dependiente (PIB) en forma de modelo de regresión. De esta forma obtenemos una interpretación más sencilla al relacionarla con un número reducido de factores y se facilita también la tarea de predicción.

6.1. Forecasting

Como vimos en (2.5) se puede obtener una predicción a 1 paso de una variable X_{it} a partir de la regresión de $X_{i,t+1}$ sobre $\hat{F}_t$ (y posibles retardos de los factores). Recordemos que el *forecast* se realiza sobre la información condensada en los factores y sobre retardos de la propia variable. Consideremos que interesa la predicción para una variable Y_t, que recordemos, puede estar contenida en X_t o no. En caso de que Y_t no esté en X_t en la ecuación de predicción entraría un término extra para introducir posibles lags de la serie, ya que al estimarse los factores sin considerar Y_t podría no estar toda la información relevante recogida en ellos. Por ejemplo no se incluye una variable de interés en el conjunto de datos cuando se trata de una medida macroeconómica, al estar conducida por un aglomerado de otras variables microeconómicas. Sea la ecuación de *forecast*:

$$Y_{t+h} = \alpha(L)f_t + \delta(L)Y_t \tag{6.1}$$

En (6.1) mostramos la expresión general donde se incluyen retardos de la variable, en caso de considerar Y_t en X_t bastaría con eliminar el último término de la ecuación.

Para realizar el forecast a h pasos se puede proceder de dos formas. La primera, directa, es realizar la regresión de Y_{t+h} sobre sus retardos y $\hat{F}_t$. Otra alternativa es un proceso iterativo, estimando primero un proceso VAR para F_t y después utilizarlo para completar sucesivas predicciones a un paso hasta llegar al horizonte h. La selección de una alternativa u otra teóricamente no está clara. La forma directa elude posibles especificaciones erróneas en el modelo VAR para Y_t y F_t, pero en caso de que esté bien especificado la predicción iterativa es más eficiente.

Mariano y Murasawa (2003), una vez estimados los factores trimestrales, modelan la relación entre el PIB y los factores como una función lineal:

$$Y_t = c + V(L) f_t^Q,$$

donde $V(L)$ actúa como filtro que pasa la información contenida en f_t^Q a los valores contemporáneos y futuros de Y_t. Cuevas y Quilis (2012) sugieren expresar $V(L)$ como un modelo Box-Jenkins expresado en forma racional:

$$Y_t = c + \frac{\omega_s(L)L^b}{\delta_r(L)} f_t^Q + \frac{\theta_q(L)}{\Phi_p(L)} u_t, \quad u_t \sim N(0, v_u),$$

donde $\omega_s(L)$, $\delta_r(L)$, $\theta_q(L)$, $\Phi_p(L)$ son polinomios de retardo de orden s, r, q, p respectivamente.

6.2. Autoregresión vectorial aumentada con factores

A lo largo del trabajo se han comentado los problemas de los modelos VAR. El primero consiste en que con N variables el número de parámetros crece con N^2, lo que hace inabordable la estimación cuando N/T es grande. El segundo es que al aplicar estos modelos considerando un vector de baja dimensión el espacio generado por las innovaciones del VAR puede no abarcar el de los shocks estructurales, es decir, con las innovaciones del modelo no se pueden generar los shocks verdaderos, lo conduce a error. Esto se conoce como el *problema de invertibilidad*.

En la práctica se ha comprobado cómo un número reducido de factores (del orden de 1 o 2 en muchos casos) son capaces de explicar las dinámicas de variables económicas de interés. Poder explicar los movimientos económicos a partir de un MFD ha motivado la investigación de nuevos modelos para incorporar la información de los factores en modelos VAR, clasicamente utilizados en este contexto.

A raíz de todo esto, se introduce un modelo VAR que puede emplearse para la identificación de los shocks verdaderos cuando contamos con un elevado número de series temporales que potencialmente pueden contener información sobre los shocks subyacentes. Esta alternativa se basa en el MFD y se

conoce como modelo VAR aumentado con factores (FAVAR, por sus siglas en inglés). Stock y Watson (2005) resuelven el modelo a partir de estimaciones de los factores y muestran el buen comportamiento del FAVAR sobre el conjunto de datos en Stock y Watson (2002a) para encontrar un modelo que ajuste los shocks en la política monetaria de EEUU y el índice de producción industrial. También realizan estudios de Monte Carlo concluyendo que el método estima el modelo con precisión.

Para la representación de este modelo interesa expresar el MFD como en (1.2), donde en la ecuación de observación se permite que entren lags del propio vector de series. El razonamiento es totalmente análogo al mostrado hasta ahora, simplemente se añade un término más en la ecuación de observación del MFD expresado como un modelo de espacio de estados. La forma estática del MFD, conservando la notación y dimensiones en (2.6), (2.7) en este caso es:

$$X_t = \Lambda F_t + D(L)X_{t-1} + \xi_t \tag{6.2}$$

$$F_t = \Phi(L)F_{t-1} + G\eta_t, \tag{6.3}$$

donde $D(L) = \text{diag}(\delta_1(L), \ldots, \delta_N(L))$, con $\delta_i(L)$ un polinomio de retardo extrapolado de (3.7). En terminología del modelo de espacio de estados, la ecuación (6.2) es la ecuación de medida y (6.3) la ecuación de estado. El término ξ_t se deriva de (3.7) y de introducir en el modelo lags del vector X_t. Cabe destacar que seguimos conservando todas las hipótesis postuladas hasta ahora sobre estacionariedad de las series, media cero y desviación típica unitaria, de donde se sigue que:

$$\Sigma_X = \Lambda\Sigma_F\Lambda' + \Sigma_e \tag{6.4}$$

Bernanke *et al.* (2005) y Stock y Watson (2005) muestran que el problema del modelo VAR puede solventarse imponiendo ciertas restricciones derivadas del MFD. La representación VAR del MFD se obtiene sustituyendo (6.3) en (6.2) y agrupando términos. La ecuación para la i-ésima serie en este modelo VAR es:

$$X_{it} = \Lambda_i\Phi(L)F_{t-1} + \delta_i(L)X_{i,t-1} + \epsilon_{X_i,t}, \tag{6.5}$$

donde $\epsilon_{X_i,t} = \Lambda_i G\eta_t + \xi_t$. Combinando (6.5) con la ecuación que modela la dinámica de los factores (6.3) se llega a la representación del MFD estático (6.2), (6.3) como un modelo VAR, es decir, el modelo FAVAR:

$$\begin{pmatrix} F_t \\ X_t \end{pmatrix} = \begin{pmatrix} \Phi(L) & 0 \\ \Lambda\Phi(L) & D(L) \end{pmatrix} \begin{pmatrix} F_{t-1} \\ X_{t-1} \end{pmatrix} + \begin{pmatrix} G & 0 \\ \Lambda G & I \end{pmatrix} \begin{pmatrix} \eta_t \\ \xi_t \end{pmatrix}, \tag{6.6}$$

Con esta representación el número de parámetros libres es $O(N+q^2)$. Además todos los parámetros se pueden estimar a partir de regresiones de X_t sobre $\hat{F}_t$ y sus residuos. Por ejemplo, dado un estimador para F_t, $\hat{F}_t$, los coeficientes del VAR se estiman a partir de una regresión de $\hat{F}_t$ sobre sus lags para

obtener el estimador $\hat{\Phi}(L)$ de $\Phi(L)$. El número de factores considerados se calcula a partir de alguno de los criterios introducidos anteriormente.

Stock y Watson (2005) proponen un proceso iterativo. En un primer paso se estiman los factores estáticos, $\hat{F}_t$, a partir del método de CP y la matriz de cargas $\hat{\Lambda}$ y de coeficientes $D(L)$ se estiman por un ajuste por mínimos cuadrados sobre $\hat{F}_t$. Finalmente se re-estiman los factores $\hat{F}_t$ por CP de $X_t - \hat{D}(L)X_{t-1}$. Se itera este proceso hasta alcanzar algún criterio de convergencia.

Capítulo 7

Paquetes en R para el Modelo Factorial Dinámico

En este capítulo se resumen las distintas funciones y técnicas de estimación disponibles en el software R para el MFD. Al ser una metodología con un desarrollo muy reciente, solo se dispone de dos paquetes para trabajar con MFDs. De hecho, uno de ellos está enfocado hacia los modelos de espacio de estados, pero dada su correspondencia con el MFD dinámico podemos aprovechar las técnicas desarrolladas para estos modelos. El otro paquete, específico para el MFD y publicado en agosto de 2019 tiene un limitado repertorio de funciones. En definitiva, uno de los principales campos de desarrollo para el MFD es profundizar en su implementación práctica en los diversos softwares estadísticos, en aras de facilitar su implementación, acercarlo a más investigadores y motivar la continuación de un mayor desarrollo del modelo.

7.1. Paquete MARSS

El paquete *MARSS* (Holmes *et al.* (2020)), desarrollado por empleados del gobierno federal de EEUU, estima los parámetros de un modelo linear MARSS (Multivariate Auto-Regressive State-Space) con errores gaussianos. A pesar de no ser el objetivo principal del paquete el ajuste de MFDs, los modelos para lo que está pensado entran dentro de los de estado de espacios, a los que como ya vimos se puede adaptar el MFD. Con este paquete podemos ajustar modelos MARSS a un conjuntos de series multivariante vía máxima verosimilitud empleando un algoritmo EM, introducido en la Sección 3.3.1. Como ya comentamos, en la literatura se sugiere la implementación no paramétrica, por lo que este paquete nos sitúa en el marco idóneo para el ajuste de MFDs. También se puede emplear un algoritmo BFGS (método quasi-Newton), que puede ser más eficientes en algunas aplicaciones. El algoritmo EM

se aproxima rápidamente a las vecindades de la máxima verosimilitud, pero tiende a tardar más para llegar a ella. Por otra parte, el EM es robusto ante las condiciones iniciales. Una práctica habitual es emplear el algoritmo EM para aproximarse a la solución y el BFGS para "pulir" la estimación.

El paquete recurre al filtro de Kalman, ya implementado en el ecosistema de R para suavizar las estimaciones de los factores tras cada iteración del algoritmo EM, en linea con lo comentado en la Sección 3.3.1. Se emplea el filtro de Kalman para obtener estimaciones de los factores latentes condicionados a $\{1, \cdots, t-1\}$ y el suavizador de Kalman para estimaciones condicionadas a $\{1, \cdots, T\}$.

Un modelo MARSS con errores gaussianos toma la forma:

$$F_t = B_t F_{t-1} + u_t + C_t c_t + G_t w_t, \quad w_t \sim MVN(0, Q_t) \tag{7.1}$$

$$X_t = Z_t F_t + a_t + D_t d_t + H_t v_t, \quad v_t \sim MVN(0, R_t) \tag{7.2}$$

$$F_1 \sim MVN(\pi, \theta) \quad o \quad F_0 \sim MVN(\pi, \theta) \tag{7.3}$$

La ecuación (7.1) es la ecuación de estado y (7.2) la de observación. Con esta formulación, los datos observables se encuentran en X. Los vectores c_t y d_t se conocen como variables exógenas o covariables. Ya se ha comentado la equivalencia entre el AFD y los modelos de estado-espacio, por lo que podremos expresar el modelo como un MARSS, el cual es más general.

Holmes *et al.* (2014) proponen en su trabajo un MFD donde los factores siguen un proceso de paseo aleatorio. En este caso, reescriben las ecuaciones (2.1), (2.2) como:

$$f_t = f_{t-1} + w_t, \quad w_t \sim MVN(0, Q) \tag{7.4}$$

$$F_t = Z f_t + a + v_t, \quad v_t \sim MVN(0, R) \tag{7.5}$$

$$F_0 \sim MVN(\pi, \theta)$$

De esta forma, tenemos el MFD expresado como un modelo MARSS fijando la matriz B en (7.1) igual a la identidad $m \times m$ y el vector $u_t = 0$. Los parámetros c_t y d_t suelen suponerse nulos en la mayoría de modelos MARSS y aquí se conserva también esa estructura. El parámetro a actúa como compensación en la ecuación y suele omitirse también salvo en casos particulares donde sea de utilidad contar con un término de esta forma. En lo que sigue, también supondremos $a = 0$.

El modelo anterior necesita una serie de restricciones para ser identificable. Holmes *et al.* (2014) sugieren que todos los elementos por encima de la diagonal de Z sean fijados a cero y $Q = I_r$. Se siguen considerando las series estandarizadas por conveniencia, pero no es necesario para estimar este tipo de modelos. El paquete MARSS permite imponer diferentes estructuras sobre los parámetros del modelo para la estimación, lo que nos dota de una herramienta para estimar la gran mayoría de modelos con los que nos encontremos. Como ya se ha comentado, los factores y las cargas no son separadamente identificables. Al aplicar el algoritmo EM al modelo tal como se ha planteado,

obtendremos una solución cualquiera de entre todas las equivalentes. Para rotar los factores y obtener la estimación óptima emplearemos la rotación *varimax*, que busca la matriz H que crea la mayor diferencia entre las cargas. Es decir, busca que las filas de Z sean más similares a $(0.8, 0.1, 0.1)$ que a $(0.2, 0.2, 0.2)$. El propio paquete MARSS devuelve la matriz H óptima por lo que solo tendremos que realizar los productos correspondientes para rotar las cargas y los factores. Notemos que existirá la matriz de rotación cuando Z tenga más de una columna (estemos en un modelo con más de un factor). Habitualmente, tras la rotación, muchas componentes de las cargas se van a cero, lo cual facilita la interpretación del modelo estimado.

El modelo MARSS nos da el marco perfecto también para introducir covariables, como en (1.2). Simplemente se suma en la ecuación (7.5) un término Dd_t, donde d_t es el vector de covariables y D la matriz del efecto de las covariables. Una restricción que encontramos es que no pueden contener valores perdidos, la cual era una de las ventajas de los métodos paramétricos. En la práctica no será un problema mayor ya que se puede realizar un ajuste de la serie previo para eliminar los valores ausentes como ya comentamos.

7.2. Paquete nowcasting

El paquete nowcasting (de Mattos *et al.* (2019)) considera el MFD en su forma dinámica como se ha introducido, considerando que los errores de los factores siguen una estructura VAR con errores normales. También se supone que las componentes idiosincráticas siguen un modelo autorregresivo como en (2.4).

En este paquete contamos con la función *Bpanel*, con la cual se puede realizar todo el proceso para transformar el panel de datos en un conjunto de series estacionarias, estandarizadas y sin la presencia de datos atípicos (siguiendo el criterio de Giannone *et al.* (2008) eliminando los datos que disten más de 4 veces el rango intercuartílico respecto a la mediana). La función incluye 8 diferenciaciones diferentes, pudiendo aplicar cada una a distintas series. Finalmente, también permite eliminar las series que contengan un cierto porcentaje de valores ausentes, el cual se puede modificar y viene por defecto como 1/3. También incluye una función para calcular los criterios de información introducidos en la Sección 4.1.2 una vez tenemos el panel de series estacionarias y estandarizadas, permitiendo aplicar los tres criterios propuestos por Bai y Ng (2002). Una vez determinado el número de factores, implementa también el criterio de información recogido en la Sección 4.2 e introducido por Bai y Ng (2007) para determinar el número de shocks de los factores.

En cuanto a la estimación del modelo ofrece dos posibilidades. La primera es la estimación en dos pasos introducida en la Sección 3.3.2, lo que amplía la opción para el ajuste del modelo respecto al paquete MARSS. Una funcionalidad interesante que nos brinda es que en caso de considerar una variable dependiente medida en una frecuencia más baja que las explicativas, la propia función transforma los factores a la frecuencia más alta antes de realizar la regresión, siguiendo la agregación de Mariano y

Murasawa (2003). Por ejemplo, si interesa el forecast para el PIB (trimestral) a partir de varias series mensuales, el paquete estima los factores como variables mensuales, procede con la agregación usada en Mariano y Murasawa (2003) para obtener factores estimados con frecuencia trimestral, y finalmente realiza la regresión del PIB sobre los factores trimestrales para ajustar la serie y realizar el forecast.

La segunda alternativa es la estimación mediante el algoritmo EM introducido en la Sección 3.3.1. Este método permite considerar un conjunto de datos con frecuencias mixtas y considerar que ciertos factores afectan solo a un determinado bloque de variables. Con este estimador no se necesita modificar las frecuencias de los factores ya que estima los valores ausentes y se realizan las predicciones a partir de las bridge equations como en (3.12) según se indica en de Valk *et al.* (2019).

El paquete devuelve los factores estimados, así como todos los parámetros relacionados con el modelo ajustado e incluso los valores estimados por el filtro de Kalman para las observaciones ausentes. Con esta información el propio paquete implementa el forecast a hasta 12 horizontes de la variable de interés. La gran limitación del paquete es que solo permite la inclusión de variables mensuales o trimestrales, y que en el forecast no incluye lags de la propia variable. En el marco económico más general no supone ninguna restricción, ya que habitualmente la variable de interés es un agregado como el PIB y las variables explicativas se miden con baja frecuencia, por lo que sin duda está orientado para este campo. Es muy cómodo e intuitivo en cuanto a que podría ajustarse un modelo y obtener una predicción con 3 funciones del propio paquete, sin necesidad prácticamente de conocer qué hay detrás de las funciones. En este trabajo se cita ya que puede ser de interés y utilidad, pero la baja flexibilidad que permite hace que en la mayoría de instituciones se decanten por el paquete MARSS.

Capítulo 8

Aplicación práctica

En este Capítulo aplicaremos todo lo introducido a lo largo del trabajo con el objetivo de construir un modelo que le permita a la entidad obtener una predicción anticipada del PIB a partir de un panel de covariables relativas a Galicia. Se presenta a continuación la implementación a principios de julio de 2020 para predecir el PIB gallego en el trimestre anterior, el actual y el próximo, i.e. para $Q2$, $Q3$ y $Q4$ [1] de 2020. Estas predicciones se toman de forma que se realizan los tres tipos de predicción introducidos en el Capítulo 1, realizando de esta forma *backcast* para $Q2$, *nowcast* en $Q3$ y *forecast* para $Q4$. Para $Q3$, a pesar de técnicamente estar realizando *nowcasting*, se podría incluir dentro del *forecasting* debido a la precocidad con la que se realiza la estimación y el hecho de que hasta el final del mes no tenemos ningún datos conocido relativo al trimestre. De esta forma, el último dato disponible para las series que se consideran es el de junio de 2020.

Para la entidad resulta de vital importancia el correcto ajuste y predicción del PIB, más aún el conocer estimaciones anticipadas de cara a la toma de decisiones y debido al considerable retraso con el que se publica su valor. Esta variable sintética nos indica cómo evoluciona la economía en un territorio y da una idea del "estado de la economía" en la región, de ahí su trascendencia. Profundizando en su definición, es una variable que mide el valor monetario total de los bienes y servicios finales producidos para el mercado, dentro de las fronteras de un territorio, en un año dado. Queda clara la importancia de este índice para la comprensión del entorno macroeconómico, ya que resume y reproduce la realidad de la circunscripción de interés. En la Figura 8.1 mostramos una gráfica de la variable de interés para indagar un poco más en su dinámica y comprender mejor los sucesivos cálculos y resultados. En la figura se resaltan los dos períodos de la historia reciente más destacados en cuanto a cambios bruscos en el contexto económico. Ambos coinciden con las dos últimas crisis, la financiera del 2008 y la última crisis generada por el virus COVID-19. Queda patente cómo el PIB refleja el estado de la economía, cayendo claramente su nivel en ambas depresiones. El último dato disponible, para $Q1$ de 2020 rompe con fuerza la dinámica ascendente previa de la serie, lo que dificultará los ajustes y predicciones. Además, puede

[1] Nos referiremos a los distintos trimestres mediante Qi, siendo $i \in \{1, 2, 3, 4\}$ el trimestre correspondiente.

apreciarse como históricamente la serie del PIB ha sido una curva muy suave, produciéndose en este último dato una caída sin precedentes. Este fenómeno también se produce en las series de covariables, las cuales también han sufrido un importante cambio de nivel sin precedentes en muchos casos. El hecho de considerar toda la información disponible para las covariables hace que el ajuste a partir de ellas sea más preciso, ya que se dispone de más información en cuanto a la continuidad y pendiente del declive económico, y no sólo un dato "atípico" al final de la serie, sin ninguna información extra acerca de su persistencia e importancia. De esta forma, de nuevo, el MFD se ajusta al contexto en que nos encontramos y aparece claramente justificado, permitiendo que se introduzca gracias al panel considerado, disponible con mayor frecuencia, información anticipada relativa a la depresión del PIB. Dado lo brusca que ha sido la caída de nivel, poder incorporar información adicional y disponible antes y con mayor frecuencia que la del propio PIB será de gran interés para obtener una predicción precisa, más teniendo en cuenta lo atípico de la situación.

En el modelo se considera la serie del PIB log-diferenciada. De esta forma eliminamos la tendencia en la variable, de forma que en la regresión se introduzca una serie estacionaria y además quede representado el PIB en la medida que realmente interesa a la entidad, que es la variación porcentual trimestre a trimestre. El nivel en sí de la serie no proporciona información relevante, pero si el cambio porcentual de un trimestre a otro. Así, al considerar la serie log-diferenciada la expresamos en cambios trimestrales porcentuales. A lo largo de este capítulo todos los ajustes y predicciones se expresan en esta escala de incremento porcentual trimestre a trimestre, por lo que por ejemplo un 5 en cualquier gráfica significa un incremento del 5 % respecto al trimestre anterior. Nos referiremos a esta escala como *escala porcentual*.

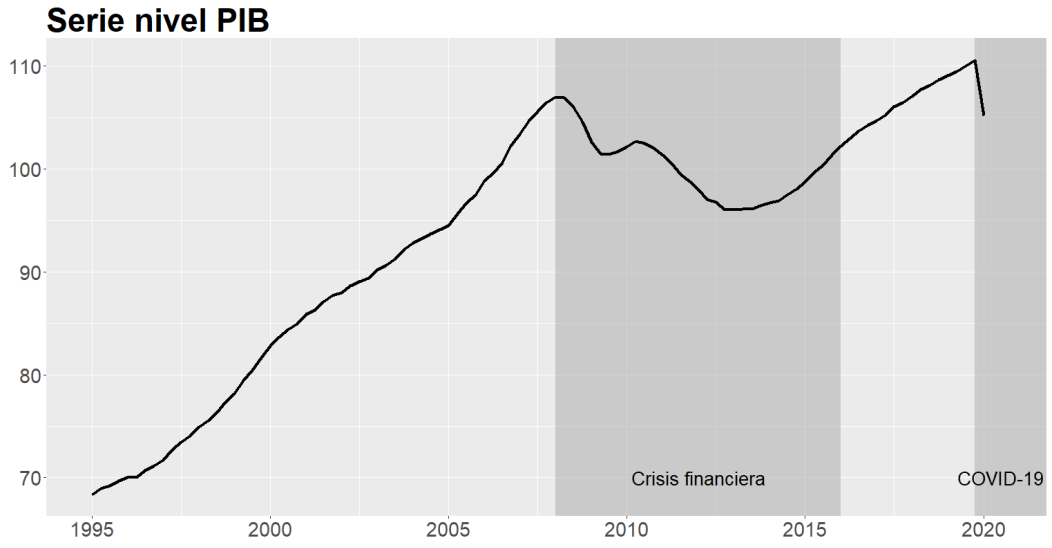

Figura 8.1: Serie del nivel del PIB considerada en el estudio como variable de interés, medida entre 1995 y el primer trimestre de 2020. Se representan sombreadas las dos últimas depresiones económicas producidas por las respectivas crisis de 2008 y 2020.

La implementación de esta parte práctica se presentará siguiendo el orden en que se han ido sucediendo los capítulos previos. En la primera parte comenzamos introduciendo las variables, con las que se comienza el estudio y se realiza todo lo relativo al tratamiento del panel de datos. En base al seguimiento económico que realiza el banco de la economía gallega se selecciona una batería de indicadores económicos que pueden resultar de interés para el ajuste y predicción del PIB, considerando variables relacionadas y/o influyentes sobre el PIB históricamente. A partir de este panel de variables, se justifica la selección de las que se considerarán en el modelo, y una vez seleccionado el panel definitivo se procede con las transformaciones pertinentes para convertirlas en estacionarias y estandarizadas como requiere el MFD. Sobre la serie del PIB, la cual consideraremos fuera del panel de datos a partir del cual estimamos el factor latente, también se realizan las transformaciones necesarias para tener una serie estacionaria. Notemos que al no incluir esta serie dentro del panel para la estimación del factor no es necesario considerarla estandarizada. Como ya se ha comentado, el verdadero interés reside en predecir la variación que sufrirá el PIB más que su valor en sí. Todas las covariables se consideran en logaritmos y se diferencian antes de ser introducidas en el modelo con dos motivos. El primero es transformarlas en series estacionarias sin presencia de tendencias, y el segundo relacionar las variables en los mismos términos que la variable de interés, i.e. en variaciones trimestrales con la misma *escala porcentual*.

A partir del panel ya formateado correctamente, se procede con las estimaciones de los distintos parámetros relativos al modelo. En primer lugar debe estimarse el número de factores latentes. Por medio de los dos tipos de criterios introducidos en el Capítulo 4 se justifica considerar el MFD con un factor latente. Una vez tenemos transformadas correctamente todas las variables y justificado que consideraremos un solo factor en el ajuste, este se estima a partir del panel de covariables seleccionadas anteriormente. Para la estimación de las cargas y factores del MFD se emplea la función *MARSS* del paquete *MARSS*, introducido en la Sección 7.1. Se estima un modelo en donde el factor sigue un proceso autorregresivo determinado por:

$$f_t = Bf_{t-1} + w_t, \quad w_t \sim N(0, Q)$$

con $B, Q \in \mathbb{R}$.

Todas las covariables están medidas con frecuencia mensual. De esta forma el factor que se ajusta tiene también frecuencia mensual, por lo que tras estimarlo, tendremos que convertirlo en una variable trimestral por medio de la transformación introducida en la Sección 3.5.

Finalmente se realiza el ajuste del PIB sobre el factor por medio de una regresión con corrección del error mediante un modelo ARIMA. Los parámetros de la regresión se estiman con todos los datos disponibles para el PIB (desde $Q1$ de 1995 a $Q4$ de 2019) sobre el factor estimado correspondiente a esos instantes. Se representa este ajuste para comprobar la coincidencia del factor estimado con el PIB, siendo capaz de captar los distintos ciclos macroeconómicos y movimientos de la serie. También se calcula el *backcast* para el $Q2$ de 2020 a partir del factor estimado, para el cual, al tener información disponible para las covariables hasta junio, se obtiene una estimación a partir de los valores reales del

panel. La otra predicción que se realiza es el *nowcast* para el $Q3$. En este caso, los valores del factor para los tres meses relativos a este trimestre se estiman por medio del MFD, ya que no se dispone de ningún dato para ajustar el factor en esos meses a partir del panel. El *nowcast* para el PIB se calcula finalmente a partir del factor estimado para el $Q3$ mediante la regresión ajustada previamente. Igualmente se procede para la estimación del *forecast* para el $Q4$.

Se concluye con un pequeño ejercicio de simulación, en donde se repiten los pasos expuestos anteriormente suponiendo que disponemos del panel en instantes pasados, tal como se tendría en tal momento teniendo en cuenta el retraso de publicación de las distintas covariables y del PIB. En este caso se pretende predecir el valor del PIB para $Q1$ de 2020, el cual estimamos realizando un MFD con el panel tal como estaría disponible entre octubre de 2019 y mayo de 2020 (fechas para las cuales aún no se dispondría del dato publicado oficialmente para el PIB dado el retraso de 50 días en su publicación). Se muestra cómo varía la predicción conforme nos acercamos al $Q1$ (y consecuentemente se conocen más datos), obteniendo predicciones que consideran e incluyen toda la información disponible día a día conforme se publica. Esta sección puede que sea la más destacada, en cuanto a que resume toda la filosofía e interés del MFD y demuestra su buen funcionamiento práctico.

8.1. Panel de datos y selección de variables

8.1.1. Serie del PIB

La serie trimestral del PIB con la que se llevará a cabo este estudio es el PIB a precios constantes o PIB real, cuya frecuencia es trimestral. Para poder analizar la evolución a lo largo del tiempo se escala esta variable en un índice con base en el año 2015. En el marco económico para esta serie se trabaja con el índice corregido de calendario y estacionalidad y con un tratamiento de consistencia transversal y temporal de forma que la serie sea consecuente en toda su historia. Esta es la serie final con la que se trabaja y es publicada por el IGE con un desfase de 50 días. Para su introducción en el modelo se considera la serie log-diferenciada como ya se ha comentado, de forma que se ajustan las variaciones porcentuales trimestrales y que la serie considerada esté en el mismo formato que las covariables del panel.

8.1.2. Panel de covariables

El conjunto de datos considerado inicialmente consiste en 24 variables referentes a los distintos sectores macroeconómicos que conforman el PIB. Todas las variables son publicadas por el IGE y se obtienen por medio de la automatización para su descarga implementada en R desde la entidad. La selección inicial de variables está basada en el conocimiento que los expertos del área de Estudios de la entidad tienen sobre las series y su relación con el PIB. Históricamente han sido relevantes en su

predicción y han demostrado tener relación con él, motivando considerarlas en un primer momento. En base a un estudio de las correlaciones de las variables con el PIB, la historia de las series disponibles y el criterio de los expertos se selecciona un subconjunto de las variables para conformar el panel en los sucesivos cálculos. Las variables consideradas finalmente en el panel se recogen en la tabla 8.1 con un "Sí" en la columna "Panel", indicando que se toman para el estudio, acompañadas del intervalo para el que se dispone de observaciones, el sector económico al que hacen referencia y una pequeña explicación del tratamiento que se ha hecho sobre cada una en caso de que así sea. Con este panel se pretende contar, con un número contenido de series, con variables heterogéneas referentes a los distintos sectores que determinan al PIB y que considere un abanico de variables relativas a los ámbitos más destacados que afectan al entorno económico y por consiguiente al PIB. Para cada componente del PIB contamos con varias variables históricamente relevantes en su modelado, de forma que en el panel se considera información significativa relativa a todas las componentes del PIB. No se considera un mayor número de series siguiendo las sugerencias de Boivin y Ng (2006). Además, debido a que la mayor parte de las variables consideradas hacen referencia a agregados y medidas macroeconómicas, no se justifica introducir variables adicionales ya que cabe esperar que su dinámica se encuentre contenida en la del agregado correspondiente. Los datos están disponibles hasta junio de 2020 (al realizar en julio las estimaciones no se dispone todavía de ninguna información relativa al tercer trimestre) y se seccionan de forma que empiecen en enero de 1995 para que coincidan con el primer dato disponible para el PIB ($Q1$ de 1995).

8.1.3. Selección de variables

Como se ha comentado en la Sección 4.3 no existe una relación directa entre el número de series consideradas en el modelo y la calidad del ajuste. Siguiendo en línea con lo que sugiere Boivin y Ng (2006), no nos centraremos sólo en el número de series si no también en la "calidad" de estas. Con calidad nos referimos a variables que tengan un histórico amplio, con pocos valores ausentes y con una elevada correlación con el PIB al ser la variable de interés. La violación de alguno de estos criterios puede compensarse con buenas características en los demás o para garantizar que todas las componentes del PIB tengan algún representante en el panel de variables. En caso de alguna variable no satisfacer ninguno de los criterios ni sustentarse en el apoyo experto no será considerada en las subsiguientes estimaciones y se eliminará del panel de covariables.

Recopilando los criterios que se emplean para introducir cada variable en el MFD, nos sustentaremos en 4 pilares. Estos serán la correlación de las covariables con el PIB, su histórico disponible, el criterio experto de cara a representar todas las componentes del PIB y las cargas asociadas a cada variable en un primer MFD ajustado considerando todas las variable. En primer lugar, como sugieren Cuevas y Quilis (2012), se introducen las series cuya correlación con el PIB sea en valor absoluto mayor que 0.4. Debido a que dependiendo de la forma en que calculamos las correlaciones hay muchas que no superan este umbral, se introducen algunas series que a pesar de tener una correlación menor, satisfagan alguno(s) de los restantes criterios. Como criterio añadido a que la variable disponga de un histórico

de datos amplios y la valoración experta, en una primera prueba se ha ajustado un MFD considerando todo el panel de covariables y se comprueba el "peso" de cada una en términos de su carga asociada sobre el factor estimado. En casos de ambigüedad, donde resulta difícil tomar o no una variable, nos apoyaremos en este criterio para decantar la decisión, de forma que se eliminarán las variables cuyas cargas sean bajas. En base a estos criterios consideraremos como panel de datos el conjunto de 17 variables indicadas con un "Sí" en la columna "Panel" en la Tabla 8.1, extraídas del panel inicial de 24 variables.

Presentados los criterios, a partir del conjunto de variables facilitado por los expertos de la entidad, comenzamos realizando un estudio de la correlación de estas con el PIB. Debido a la diferencia de frecuencia entre las covariables (mensuales) y el PIB (serie trimestral) comenzamos agregando las variables como la media de los meses correspondientes a cada trimestre. Con las covariables representadas como series trimestrales, calculamos tres correlaciones diferentes, de forma que los resultados nos den una idea clara de qué series realmente tienen correlación con el PIB y en qué sentido. Las distintas correlaciones se calculan sobre las covariables del panel y el PIB de forma contemporánea, dado que esta es la relación que nos interesa y que es lo que busca ajustar el MFD. En la siguiente sección se deben convertir todas las variables en series estacionarias (por medio de diferenciaciones) para poder ser empleadas en el MFD, por lo que las correlaciones de interés son sobre las diferencias de las variables (tanto la del PIB como la de las covariables). Estas diferencias se consideran sobre el logaritmo de las series, ya que como comentamos es la transformación que consideramos para obtener un ajuste en *escala porcentual*. Se calculan correlaciones contemporáneas entre el logaritmo del PIB y el logaritmo de las distintas covariables, con 3 tipos de medida:

1. **En variaciones trimestrales:** Correlación contemporánea de la diferencia entre un trimestre y el anterior entre el logaritmo de las covariables y el logaritmo del PIB.
2. **En variaciones interanuales:** Correlación contemporánea de la diferencia entre un trimestre y el mismo trimestre del año anterior del logaritmo de las covariables y el logaritmo del PIB..
3. **En variaciones anuales:** Correlación contemporánea de la diferencia entre la media anual de las diferencias anteriores (variaciones interanuales) año a año entre el logaritmo de las covariables y el logaritmo del PIB..

Con estas tres correlaciones se tiene una idea de cómo afecta y en qué temporalidad cada variable al PIB. Como se puede ver en la Figura 8.1, la serie del PIB es suave con tendencias muy marcadas en general. Debido a esto, las correlaciones más elevadas se tienen en diferencias anuales (series más suaves), ya que el mayor ruido de las covariables hace que sea más difícil percibir sus relaciones con el PIB. En la Figura 8.2 se recogen las correlaciones consideradas para todas las covariables. En la gráfica se incluyen dos líneas horizontales a altura ± 0.4, en base al criterio que sugieren Cuevas y Quilis (2012). Dado el limitado número de series que superan este umbral, se recurre al criterio experto y a la historia disponible para incluir más de las que así se incluirían. Este gráfico se pretende que sirve como selector cualitativo, especialmente para las covariables que se criban para las posteriores estimaciones.

Nombre	Código	Sector	Año primer dato	Panel	Tratamiento serie original y notas
Consumo gasóleo+gasolina	COMB	Consumo	2005	Sí	Corregida de estacionalidad, calendario y atípicos aditivos
Matriculación de vehículos	MATVEH	Consumo	1980	Sí	Corregida de estacionalidad, calendario y atípicos aditivos
Índice de comercio al por menor a precios constantes	IXERCOM	Consumo	1995	Sí	Deflactada con base 2015
Afiliaciones a la SS	AFSS	Mercado laboral	1990	Sí	Corregida de estacionalidad, calendario y atípicos aditivos
Paro registrado	PARO	Mercado laboral	1996	Sí	
Edificación nueva planta	EDIF	Construcción	1990	Sí	Corregida de estacionalidad, calendario y atípicos aditivos
Compraventa de viviendas	COVEVIV	Construcción	2006	Sí	
Licitación oficial	LICIT	Construcción	1990	No	
Vivendas visadas	VIVIS	Construcción	2000	No	
Índice de entrada de pedidos en la industria	IENTPED	Industria	2010	Sí	Deflactada con base2015
Índice de cifra de negocios en la industria	ICNEG	Industria	2010	Sí	Deflactada con base 2015
Índice de Producion Industrial	IXERIND	Industria	2000	Sí	Deflactada con base 2015 y corregida de estacionalidad, calendario y atípicos aditivos
Industria manufacturera	MANUF	Industria	2001	Sí	Deflactada con base 2015
Producción total de vehículos	PRODVEH	Industria	1994	No	Corregida de estacionalidad, calendario y atípicos aditivos
Nivel cartera de pedidos	NCARPED	Industria	1994	No	Corregida de estacionalidad, calendario y atípicos aditivos
Nivel stocks productos terminados	NSTOPROD	Industria	1994	No	Corregida de estacionalidad, calendario y atípicos aditivos
Tendencia en la producción	TENDPROD	Industria	2003	Sí	Corregida de estacionalidad, calendario y atípicos aditivos
Transporte marítimo de mercancías	TRANSMAR	Servicios	2000	Sí	Corregida de estacionalidad, calendario y atípicos aditivos
Índice cifra de negocios en servicios	IXERSERV	Servicios	2000	No	Deflactado por el IPC
Pernoctaciones	PERNO	Servicios	2005	Sí	Corregida de estacionalidad y calendario
Viajeros entrados	VIAENT	Servicios	1980	Sí	Corregida de estacionalidad, calendario y atípicos aditivos
Transporte aéreo de pasajeros	TRANSAER	Servicios	1980	No	Corregida de estacionalidad, calendario y atípicos aditivos
Importaciones de bienes	IMPORT	Sector exterior	1995	Sí	Deflactada por el IVU de importaciones y corregida de estacionalidad
Exportaciones de bienes	EXPORT	Sector exterior	1995	Sí	Deflactada por el IVU de exportaciones y corregida de estacionalidad

Tabla 8.1: Tabla resumen del panel original de covariables propuesto por ABANCA

A continuación presentamos los sectores económicos donde se engloba cada una de las variables y explicamos el proceso por el que se han seleccionado o no las distintas variables para el panel de covariables que emplearemos en el MFD. Dentro de cada sector, damos una pequeña intuición acerca

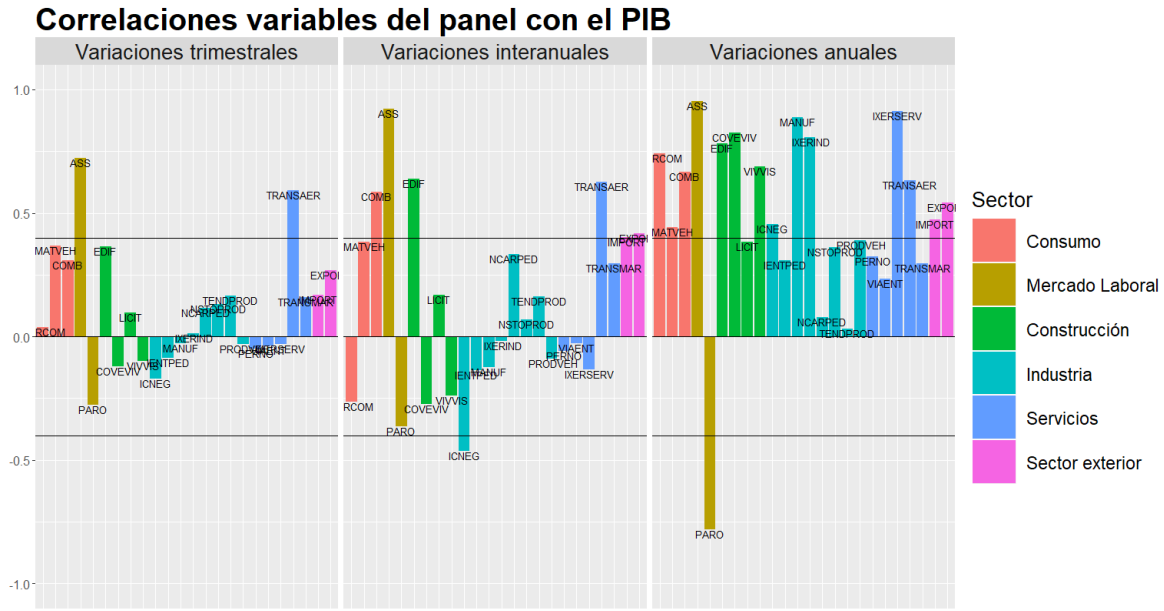

Figura 8.2: Tres tipos de correlaciones consideradas en esta memoria para cada covariable con el PIB. Por colores se destacan los distintos sectores y con las líneas verticales ± 0.4 el umbral considerado por Cuevas y Quilis (2012) para la inclusión de variables en el panel.

de la motivación de considerarlo y se justifica la selección o no de las distintas variables del panel inicial. Los razonamientos se basan en los criterios del estudio de las correlaciones, la historia disponible para la serie, las cargas estimadas con el MFD ajustado con todo el panel y el criterio experto ya comentados. Las correlaciones se resumen en la Figura 8.2, las temporalidades se pueden encontrar en la Tabla 8.1 y el valor de las cargas y/o el criterio experto se expone cuando resulta concluyente.

Consumo Este es uno de los sectores que históricamente se han considerado espejo del PIB. Es muy intuitiva la relación entre consumo (donde entran gastos personales en bienes y servicios y también el gasto realizado por el gobierno) y el PIB. En Galicia se estima que este sector representa al rededor del 70 % del PIB. Por tanto, el consumo resulta clave a la hora de analizar el comportamiento del PIB, y por esta razón, a pesar de no satisfacer los criterios de correlación (véase la Figura 8.2) consideraremos las tres variables relativas a este sector.

Mercado laboral En este sector nos encontramos con indicadores muy relacionados con la situación económica y que tienen muy poco retraso en la publicación, sobre todo la media de afiliados y el paro registrado en cada mes. Esta característica tan ventajosa jugará a favor de estas variables de cara a ser consideradas en el panel, ya que van a permitir anticipar el comportamiento de la economía. Estos indicadores se toman corregidos de calendario y estacionalidad. Dentro de este sector se encuentran las

afiliaciones a la seguridad social (SS) y el paro. En la Figura 8.2 se observa la alta correlación entre estas variables y el PIB, lo que justifica la inclusión de ambas. Las afiliaciones a la SS superan claramente el umbral de 0.4 en todas las correlaciones y el paro, a pesar de superarlo sólo en el tercer tipo de correlación, está muy próximo a el en las otras dos. Son dos de las variables económicas más conocidas y el criterio experto coincide con la intuición general sobre la influencia en el entorno económico de estas variables, concluyendo en la introducción de ambas.

Construcción El peso de la construcción dentro del PIB gallego desde el punto de vista de la oferta del Valor Agregado Bruto es del $7,7\,\%$ (excluidos impuestos). Este sector perdió parte de su peso sobre el PIB a raíz de la crisis económica del 2008, pero recientemente ha recuperado su peso y cabe esperar que influya en la dinámica de la variable de interés. Dentro de este sector se elimina la serie de *licitaciones (LICIT)* debido a la baja correlación que muestra con el PIB (la menor dentro de su sector además) y a la ausencia de justificación por ningún otro criterio. Asimismo, la carga asociada a esta variable cuando se estima el factor con el panel completo era muy baja, casi 0, por lo que se opta por no considerarla. Igualmente ocurre con la variable *viviendas visadas (VIVIV)*, la cual además cuenta con datos solo desde el 2000. El resto de variables de este sector serían susceptibles de no ser tomadas en el panel teniendo sólo en cuenta la Figura 8.2, pero teniendo en cuenta el impacto que la dinamización de este sector suele tener en la economía, se incluyen las variables *Edificación nueva planta (EDIF)* y *Compraventa de viviendas (COVIV)*.

Industria El peso del sector industrial en el PIB de Galicia se estima que es del $17,2\,\%$. Las series que se emplean dentro de ABANCA para hacer el seguimiento del sector industrial y que se han seleccionado como candidatas a este estudio son todas menos *Nivel stock productos terminados (NSTPROD)* y *Tendencia en la producción (TENDPROD)*. Además de la baja correlación de estas variables con el PIB, su estrecha relación con la variable *Nivel cartera pedidos (NCARPED)* hace que sean descartadas para el panel, con el fin de no introducir en el modelo información redundante. Tras implementar el MFD con el panel completo se obtiene una carga muy baja para *NCARPED* también, por lo que se concluye con no considerarla en el panel tampoco. El resto de variables de este sector han sido históricamente muy relevantes de cara a predecir el PIB, por lo que se recomienda su inclusión. Más adelante, cuando representemos las cargas, puede verse la importancia destacada de las variables tomadas de este sector.

Servicios El peso del sector servicios dentro de la componente de oferta del PIB de Galicia es del $69,9\,\%$ (excluidos impuestos). De entre los indicadores candidatos a entrar en el panel eliminamos la variable *Transportes marítimos (TRANSMAR)* debido a su baja correlación con el PIB (Figura 8.2). Existe una estrecha relación entre la variable *Pernoctaciones (PERNO)* y *Viajeros entrados (VIAEN)*, por lo que buscando evitar introducir información duplicada o redundante eliminamos la variable *VIAEN* por ser la que más baja correlación muestra con el PIB de las dos.

Sector exterior Debido a la inflación y las consecuentes variaciones en los precios, no podemos considerar desde un principio las series tal como están medidas, ya que año a año varía la "unidad" de medida. Deflactar consiste en el proceso de eliminar estas variaciones en la unidad de medida de forma que todos los datos históricos estén expresados sobre el mismo valor, lo cual se lleva a cabo por medio del IVU. Desde el punto de vista de la demanda, el PIB se descompone en Consumo, formación bruta de capital (FBC) y Sector exterior. Es por ello que incluiremos en el modelo las series mensuales de exportaciones e importaciones de bienes. Como el PIB está a precios constantes, las series brutas se deflactan por el índice de valor unitario (IVU) y a continuación se corrigen de calendario y estacionalidad. En cuanto a los criterios para la selección de variables, en la Figura 8.2 se aprecia una correlación elevada de las dos variables. Ambas series están disponibles desde el 1995, la misma fecha donde comienza nuestra serie del PIB. Además, en la Figura 8.3 se puede percibir como la dinámica de ambas series es muy similar a la de la serie del PIB, representada en la Figura 8.1. Sumado todo lo anterior concluimos en la inclusión de ambas variables en el panel.

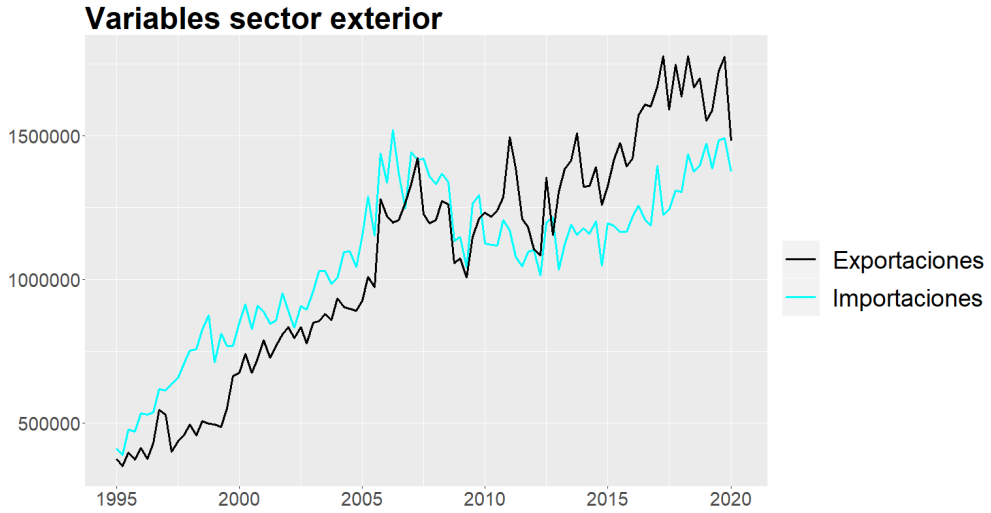

Figura 8.3: Variables relativas al sector exterior.

Transformación de las variables Una vez seleccionado el panel de variables final, indicado en la Tabla 8.1, lo transformamos en un conjunto de series estacionarias y estandarizadas. Realizaremos las mismas transformaciones tanto a las covariables como a la serie del PIB. En primer lugar consideramos la primera diferencia del logaritmo de cada serie. De esta forma se consideran los crecimientos porcentuales trimestre a trimestre (medida en la que se acostumbra a expresar el PIB) y eliminamos la presencia de tendencias. Con el Test de Dickey Fuller estudiamos la estacionariedad de las variables y comprobamos que se adecuan a las hipótesis del modelo. Las variables *Afiliaciones a la SS* y *Paro* no superan con una diferencia el test de Dickey-Fuller, por lo que les aplicamos una segunda diferencia para eliminar su tendencia. Considerando para estas dos variables el logaritmo de la serie con dos diferenciaciones obtenemos finalmente un panel de series estacionarias con un nivel de confianza $\alpha = 0.01$.

Una vez eliminada la tendencia de las series se estandarizan para que todas estén en la misma escala (i.e. que todas tengan media 0 y desviación típica unitaria), de forma que ninguna influya en demasía en el MFD debido a su escala y solo lo hagan consecuencia de su dinámica.

8.2. Estimación del número de factores

Para la estimación del número de factores latentes empleamos los dos criterios introducidos en la Sección 4.1. Se representa el scree-plot en la Figura 8.4, el cual sugiere que con 1 factor podría ajustarse una parte importante de las dinámicas del panel, además el codo más destacado parece darse considerando un solo factor. A partir del 2º factor comienza a producirse un aumento ínfimo conforme se toman más factores. Como ya se comentó, la laxitud en la interpretación es el problema principal cuando se emplea este criterio. El hecho de tomar pocas series (dentro de lo que se refiere a la filosofía del MFD) y que sean muy ruidosas dificulta la elección del número de factores. Los criterios de información de Bai y Ng (2002), como ya se indicó, constituyen una cota superior para el verdadero número de factores latentes. Debido al elevado ruido en el panel y a la fuerte estructura autorregresiva de alguna de las variables, cabe esperar que nos encontremos ante un caso en donde el número verdadero de factores sea menor al que estiman estos criterios. Además, las correlaciones reducidas que se observan en la Figura 8.2 implican una dificultad añadida a la hora de estimar el número de factores. Tras estimarlos por medio de los criterios propuestos en Bai y Ng (2002) obtenemos números demasiado elevados como para ser aplicados en un modelo intuitivo como el que se busca. Dada la falta de consistencia para la selección del número de factores, teniendo en cuenta la sugestión de considerar 1 factor latente que se obtiene del gráfico de sedimentación de la Figura 8.4, y pretendiendo un modelo sencillo y fácilmente interpretable consideramos un MFD con 1 factor.

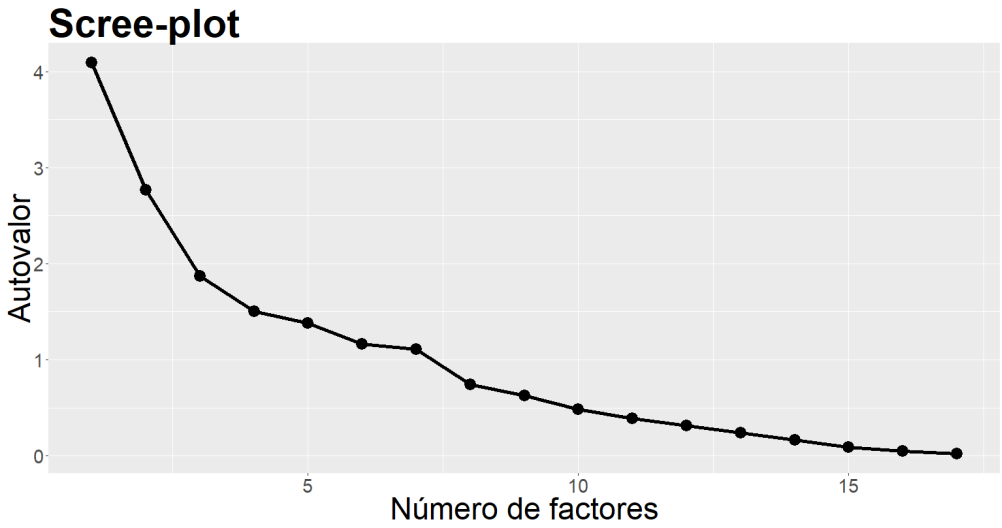

Figura 8.4: Scree-plot del panel final de covariables

8.3. Implementación del Modelo Factorial Dinámico

Una vez determinado el número de factores latentes, procedemos a estimar el factor y las cargas correspondientes a cada covariable. Con la función *MARSS* del paquete *MARSS* realizamos el ajuste. En esta implementación, el MFD expresado en términos de las ecuaciones (7.1), (7.2) y (7.3) toma la forma:

$$f_t = Bf_{t-1} + w_t, \quad w_t \sim N(0, Q)$$
$$X_t = Zf_t + v_t, \quad v_t \sim N(0, R)$$

En nuestro estudio se considera un MFD con un solo factor, por lo que $f_t \in \mathbb{R}$ y en las expresiones anteriores los parámetros $B, Q, R \in \mathbb{R}$. Al considerar 17 series en el panel y un factor en el modelo, $Z \in \mathcal{M}_{17\times 1}$. Una vez estimado el factor y las cargas, en caso de haber considerado más de un factor realizaríamos la rotación *varimax* introducida en la sección 7.1, de forma que las cargas sean lo más distintas posible dos a dos. Los factores estimados se rotarían consecuentemente a partir de la matriz de rotación empleada para las cargas, de forma que los factores rotados concuerden con el ajuste realizado sobre las variables del panel. En este estudio al considerar un solo factor no existe rotación posible y no se efectúa.

Cabe destacar también que el factor estimado es una serie estacionaria al igual que todas las del panel por construcción. Dado que las covariables del panel se introducen en frecuencia mensual, el factor ajustado también tiene la misma frecuencia. Interesa emplear el factor como variable latente que explique el PIB, por lo que debemos convertirlo en una serie trimestral. Esto lo realizamos a partir de la transformación de Mariano y Murasawa (2003), introducida en la sección 3.5. De esta forma se obtiene el factor trimestral, representado en la Figura 8.5. En esta gráfica se incluye superpuesta la serie del PIB log-diferenciada y estandarizada sobre el factor estandarizado, de forma que se puedan interpretar en la misma escala. En la gráfica puede apreciarse la dinámica similar entre ambas series, especialmente en momentos de volatilidad y movimientos bruscos en las diferencias del PIB. Aquí todavía no se ha hecho ningún ajuste del PIB sobre el factor, pero al estar en la misma escala ambas variables ya se pueden apreciar los movimientos síncronos. Más adelante, cuando obtengamos la regresión del PIB sobre el factor se visualizará mejor que tienen unas dinámicas muy similares salvo escala (el coeficiente de regresión).

Puede notarse en la representación cómo el factor captura los dos descensos más claros en la economía reciente, correspondientes a la crisis financiera del 2008 y la última crisis fruto del virus COVID-19 (ambas marcadas con una zona sombreada). Se comprueba que el factor estimado es capaz de captar los movimientos principales de las covariables consideradas y consecuentemente del PIB. El dato extremo que se produce en el $Q1$ de 2020 resulta muy difícil de ajustar debido a su brusquedad y a lo repentino que es, pero gracias a considerar la información del panel en el modelo el factor puede adaptarse mejor a este shock. Cabe destacar que los meses de enero y febrero fueron meses "normales", en cuanto a que aún no se había extendido el virus y no se tomaron medidas hasta el mes de marzo. Esto hace que el modelo tenga que ajustar este dato extremo a partir solo de la información de marzo,

lo cual lleva a una estimación conservadora para este dato. Conforme la crisis empeora y se extiende durante el $Q2$ de 2020, las covariables reflejan en sus niveles la depresión económica consecuente y el *nowcast* calculado para el $Q2$ se ajusta mejor a la situación real al incorporar más información relativa a la crisis.

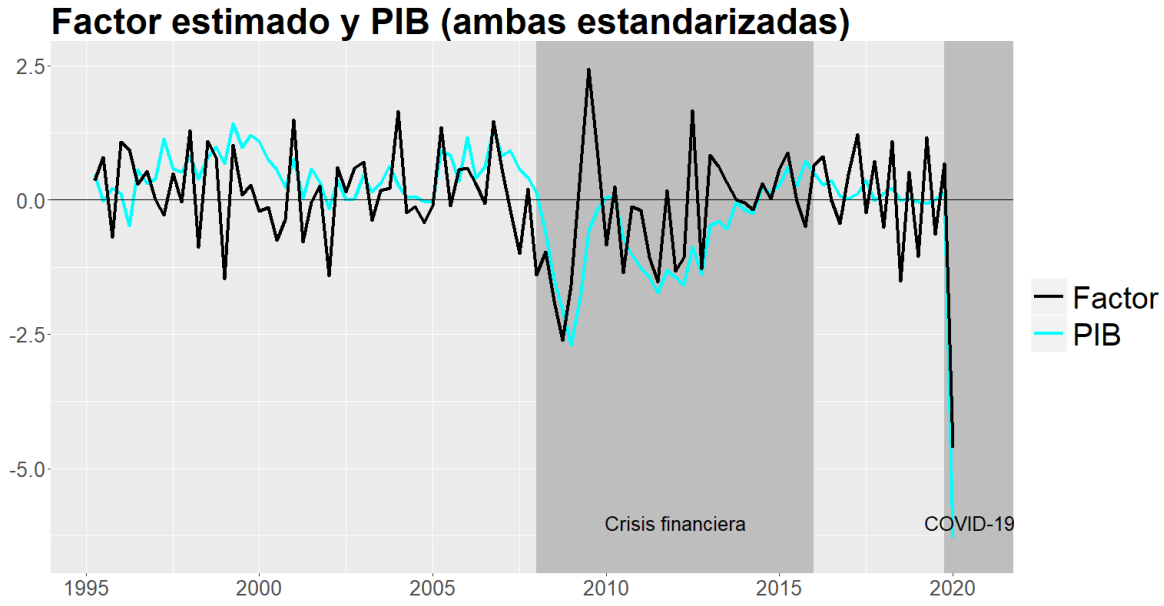

Figura 8.5: Factor estimado estandarizado y variación porcentual del PIB estandarizada, de forma que ambas series estén en la misma escala para su comparación. En las zonas sombreadas se representan los dos últimos períodos de crisis económica.

Como ya se ha comentado, las cargas asociadas a cada covariable constituyen un parámetro destacado de cara a interpretar el factor estimado, dándonos una idea de qué covariables tienen un mayor peso en el ajuste del factor y por consiguiente sobre el panel. Puede emplearse esta información para refinar las variables consideradas en el modelo o para realizar modelos sucesivos considerando solo las variables con una carga elevada. Como se ha comentado, en este estudio se realiza una primera implementación del modelo considerando el panel completo recogido en la Tabla 8.1. En este ajuste algunos índices estaban asociados con cargas muy próximas a cero, lo cual ha motivado la eliminación de algunos. En la Figura 8.6 se recogen las cargas asociadas a cada una de las covariables del panel final considerado. Las líneas horizontales indican $\pm$ la media de las cargas en valor absoluto, como referencia para estudiar qué variables son más influyentes en la construcción del factor.

Lo primero que desataca de la Figura 8.6 es que la única covariable con una carga negativa es la asociada al *Paro*, lo cual concuerda con la intuición. El resto de variables fueron tomadas en un principio como índices con una relación directa con el estado de la economía, de los que se espera que

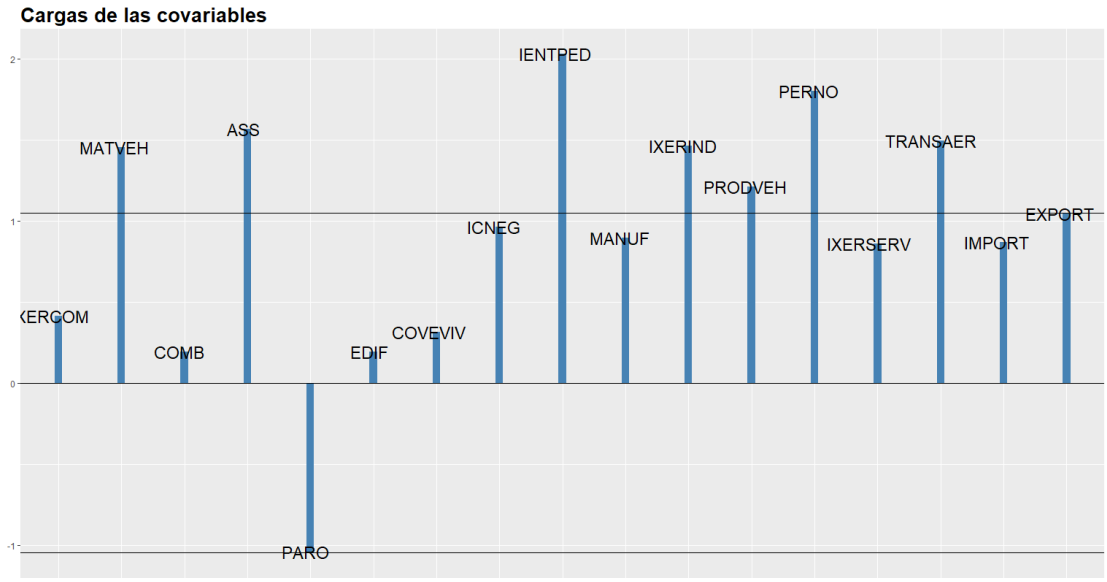

Figura 8.6: Cargas asociadas a cada covariable. La abreviatura relativa a cada variable se puede encontrar en la Tabla 8.1.

fluctuen en la misma dirección que el PIB. Puede observarse cómo las variables asociadas al sector *Servicios* tienen una elevada carga, reflejando el peso que este sector tiene en el PIB de Galicia, tal y como se ha comentado anteriormente. Como cabría esperar a priori, el sector *Industria* tiene un importante peso también dentro del estado de la economía de una región, así puede verse que salvo la variable *Índice de comercio al por menor* todas las relativas a este sector tienen una carga alta. El coeficiente que más llama la atención es el asociado a la variable *Consumo de gasóleo + gasolina* debido a su alto valor. Teniendo en cuenta la importancia del consumo dentro del PIB y el importante gasto que supone para la población en general el consumo de combustibles dentro de su presupuesto cabría esperar que fuese más determinante. Si observamos la serie relativa a esta variable, se comprueba que en los últimos años su tendencia ha sido diferente a la del PIB (la aparición de tecnologías alternativas podría estar causando este cambio de tendencia) a pesar de tener una dinámica similar hasta 2012 aproximadamente. Esto es lo que conduce a una baja carga relativa a esta variable.

8.4. Ajuste del PIB a partir del factor estimado

El primer paso para ajustar el PIB en función del factor latente es expresar ambas series en la misma frecuencia. Para ello, convertimos el factor mensual en una variable trimestral por medio de la fórmula propuesta por Mariano y Murasawa (2003) e introducida en la Sección 3.5. Con el factor representado como serie trimestral, realizamos la regresión del PIB sobre él. Se estima un modelo de regresión lineal con corrección del error corregido mediante un modelo ARIMA$(0, 1, 1)$, con el orden estimado por el criterio BIC y parámetros estimados por máxima verosimilitud.

Los parámetros del ajuste, en términos de la expresión (1.2) del PIB sobre el factor estimado se recogen en la Tabla 8.2, donde AR_1 se refiere al coeficiente del ARIMA y σ^2 la varianza del error del modelo ARIMA.

δ	AR_1	σ^2
0.19	-0.28	$3 \cdot 10^{-5}$

Tabla 8.2: Parámetros estimados para la regresión del PIB sobre el factor estimado, con corrección del error mediante un proceso ARIMA$(0, 1, 1)$.

Nótese que al disponer de datos para las covariables hasta el mes de junio, es posible realizar el *backcast* del PIB a partir del factor estimado con observaciones del panel. De igual forma, el modelo estima los valores del factor para los meses del $Q3$ y $Q4$, para los cuales aún no se dispone de ningún dato, y se realiza el *forecast* del PIB para estos trimestres a partir del *forecast* para el factor estimado y la regresión ajustada anteriormente. Cabe destacar que en este sentido podría considerarse el ajuste del PIB por medio del factor como una componente común. La única diferencia es que la "carga" asociada al PIB no se estima por medio del MFD sino con un modelo de regresión, pero obviando esto, la interpretación y filosofía es totalmente extrapolable.

En la Figura 8.7 se muestra el ajuste final que se obtiene para la serie de la log-diferencia del PIB representada en *escala porcentual* a partir del factor estimado por medio de la regresión. En esta gráfica se representan las diferencias porcentuales trimestre a trimestre, el valor que interesa seguir para evaluar el entorno económico en Galicia actualmente y a futuro. Se incluyen destacadas la estimación para el *backcast* ($Q2$), el *nowcast* ($Q3$) y el *forecast* ($Q4$). Se puede observar el buen ajuste que produce el factor, captando las dinámicas principales del PIB y también movimientos más breves. La componente común ajustada al PIB es más rugosa que la del propio PIB, debido a la mayor volatilidad de las covariables del panel. El ajuste para el último dato disponible del PIB ($Q1$ de 2020) es el único que no se aproxima correctamente al valor real. Como ya se ha comentado, el repentino y brusco movimiento de este dato, sumado al hecho de producirse en la última parte del trimestre hacen que sea imposible para el modelo capturarlo. Para el $Q2$, al contar con mayor información relativa a fechas de la propia crisis, se obtiene un ajuste mucho más razonable. En este dato se recoge no solo un mes relativo a la depresión económica, lo que hace que se ajuste mejor al claro descenso en la economía producido este trimestre. En cuanto al *nowcast*, obtenemos una predicción menos pesimista comparada con el dato anterior. Se estima para el $Q3$ una caída del nivel del PIB cercano al $1\,\%$ respecto al $Q2$, lo cual a pesar de ser un dato negativo muestra una desaceleración en la caída del PIB. Esto se debe principalmente a la apertura e incentivo de la economía a lo largo de la segunda mitad de $Q2$ y del $Q3$. Para la predicción de $Q4$ se obtiene una nueva bajada importante, de un $2.5\,\%$ respecto a $Q3$. Para este trimestre el modelo estima una nueva aceleración en la depresión del nivel del PIB, la cual resulta preocupante teniendo en cuenta que salvo por el descenso en la serie producido por el COVID-19 la predicción para $Q4$ es el valor históricamente más bajo. Probablemente la volatilidad producida esté

afectando al *forecast*, el cual es susceptible de variar conforme nos acerquemos a esa fecha y se disponga de más datos. Las series de covariables (log-diferenciadas), a pesar de seguir mostrando datos negativos, han tomado niveles más cercanos a los previos a la crisis, produciendo una "V" en sus gráficas que se refleja en el factor estimado. Al realizar las estimaciones para $Q3$ y $Q4$ sin ninguna información relativa al trimestre, cabe esperar que conforme se vayan publicando nuevos datos e introduciendo en el modelo el *nowcast* y *forecast* para el PIB se vayan aproximando al valor real. Este comportamiento se demuestra en la siguiente sección por medio de un ejercicio de simulación. Igualmente, se espera que se actualicen las predicciones cuando se conozca el dato para el $Q2$ del PIB, el cual probablemente cambie la "carga" asociada al factor del PIB dado lo extremo de sus niveles.

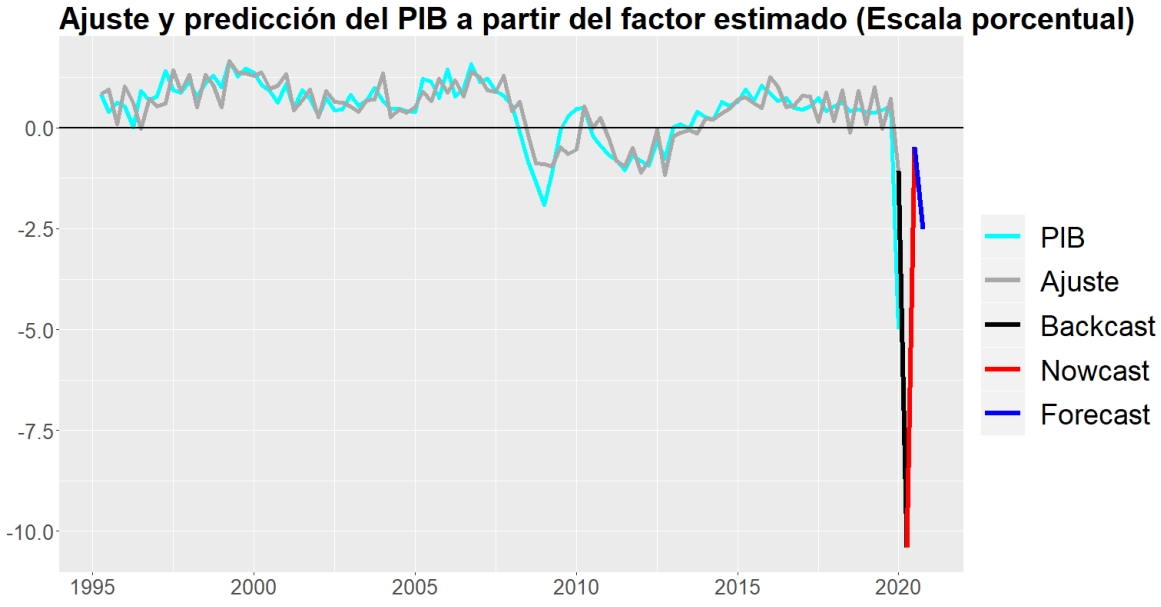

Figura 8.7: Ajuste del PIB obtenido por medio de la regresión del PIB sobre el factor estimado y distintas predicciones para el PIB a partir del *forecast* para el factor por medio del MFD.

Cálculo del Intervalo de Confianza

Finalmente calculamos los intervalos de confianza como se indica en la Sección 5. Solo se dispone de teoría asintótica para las componentes comunes y los factores en este modelo. Dentro de nuestro escenario, donde se emplea el factor para una subsiguiente regresión debemos buscar alguna aproximación de para estimar los intervalos de confianza. La propuesta de este trabajo es, apoyados en los intervalos de confianza asintóticos disponibles para el factor estimado empleando la teoría expuesta en la Sección 5, calcular el intervalo de confianza para la serie del PIB ajustada. La propuesta para dichos intervalos de confianza consiste en:

- Estimar el intervalo de confianza para el factor estimado a partir del panel de covariables, del Teorema 6 y la aproximación 5.1. Con esta aproximación por medio de una normal de la distribución del factor calculamos el intervalo de confianza.
- Dado que el factor se estima con frecuencia mensual, los intervalos de confianza se estiman en la misma frecuencia. Dado que el ajuste para el PIB es una serie trimestral, se agrega el intervalo de confianza superior e inferior de igual forma que el factor, empleando la fórmula (3.12).
- Una vez calculado el intervalo de confianza para el factor trimestral, se estima el intervalo de confianza para el ajuste del PIB como las series que se obtienen aplicando el modelo de regresión estimado previamente a los intervalos de confianza del factor.

Implementando lo anterior sobre nuestro estudio, obtenemos los intervalos de confianza con un nivel $\alpha = 0.05$ representados en la Figura 8.8. Se observa cómo la volatilidad en las covariables se traslada al factor estimado y consecuentemente al intervalo de confianza. El intervalo de confianza da una idea de la volatilidad en cada momento, siendo más amplio en momentos de volatilidad y más estrecho en períodos más "tranquilos". De esta forma a parte de poder interpretarlos de la forma habitual (el intervalo donde con una confianza del $100 \cdot (1-\alpha)\,\%$ se encuentra la serie) se pueden considerar como una medida de la volatilidad. Cabe destacar que salvo para el dato de $Q1$ de 2020 este intervalo de confianza ha contenido a la serie de interés. Será interesante conocer el dato para $Q2$ de cara a testar la fiabilidad de estos intervalos en momentos de extrema volatilidad.

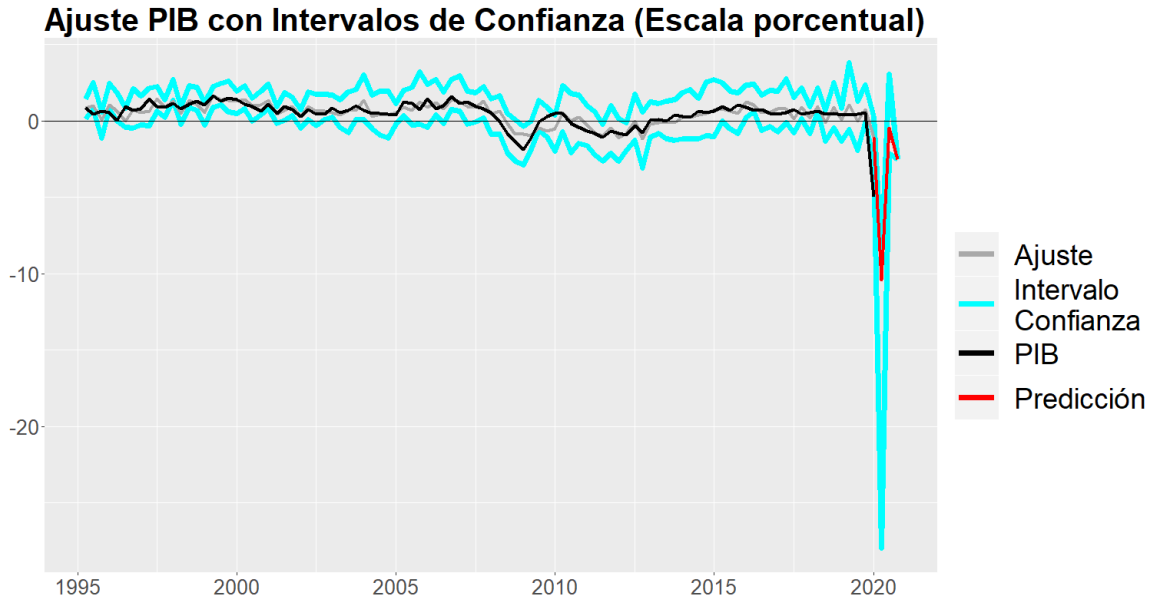

Figura 8.8: Intervalos de confianza para el ajuste y predicciones del PIB obtenidos a partir del MFD, empleando la estimación indicada en el apartado 8.4.

8.5. Predicción en tiempo real

Finalmente, se recoge una de las principales utilidades del MFD en el gráfico 8.9. En esta figura se muestra la evolución de la predicción para el PIB del $Q1$ de 2020 efectuada entre octubre de 2019 y mayo de 2020 semana a semana considerando el panel tal y como estaría disponible en cada instante teniendo en cuenta los retrasos de publicación de cada variable. Los retrasos de publicación de cada una de las covariables se recogen en la Tabla 8.5. De esta forma notemos que de nuevo realizaremos los tres tipos de predicción introducidos. En la Figura 8.9 también se recoge la predicción obtenida a partir del modelo de referencia, el modelo ARMA. El orden del modelo se obtiene a partir del criterio BIC y se estiman por máxima verosimilitud los parámetros en cada uno de los instantes. A priori este modelo tiene varios inconvenientes respecto al MFD. En primer lugar solo tiene en cuenta la propia serie del PIB y no incorpora información externa en la cual apoyarse para obtener una mejor predicción. Otro inconveniente es que se basa exclusivamente en la propia serie, lo que implica que se obtiene la misma estimación para todo el trimestre. Considerar estimaciones que no incorporen toda la información disponible con mayor frecuencia resulta poco justificable en la práctica. En esta sección ajustamos para cada uno de los paneles considerados en cada instante un modelo ARMA para la predicción del $Q1$ de 2020. Lo primero que se hace notar es que se obtiene la misma estimación en todos los instantes considerados al basarse solo en la información disponible para el PIB, algo que de nuevo no resulta de interés. Además la predicción del modelo ARMA, al no contar de información externa, predice para el $Q1$ un valor prácticamente idéntico al del $Q4$ de 2020.

Variable	ASS	PARO	IMPORT	EXPORT	EDIF	COVEVIV	ICNEG	IENTPED	MANUF	IXERIND	PRODVEH	IXERCOM	MATVEH	COMB	PERNO	IXERSERV	TRANSAER
Retraso (en días)	3	7	50	50	80	70	10	120	30	38	90	29	30	10	51	51	51

Tabla 8.3: Retrasos en la publicación de cada una de las variables consideradas, en días

Dada la filosofía detrás del MFD cabe esperar que conforme se vayan publicando más datos se disponga de más información y se produzca una consecuente mejora en el ajuste del PIB. La gráfica deja claro cómo la predicción para el PIB va rectificando hacia el verdadero valor conforme se conocen más datos. El modelo no es capaz de captar el dato extremo que se publica para $Q1$ de 2020, pero podemos observar cómo conforme se publican datos relativos al $Q1$ esta información se incluye en el MFD y la predicción se aproxima mejor al valor verdadero. Claramente las predicciones obtenidos con el MFD mejoran las del modelo de referencia. En todo momento la predicción para el $Q1$ es mejor con el MFD y evoluciona hacia el valor verdadero conforme nos acercamos a la fechas de publicación de $Q1$. En base a todo lo comentado y al ejemplo representado en la Figura 8.9 concluimos que el MFD mejora claramente las predicciones y filosofía del modelo de referencia.

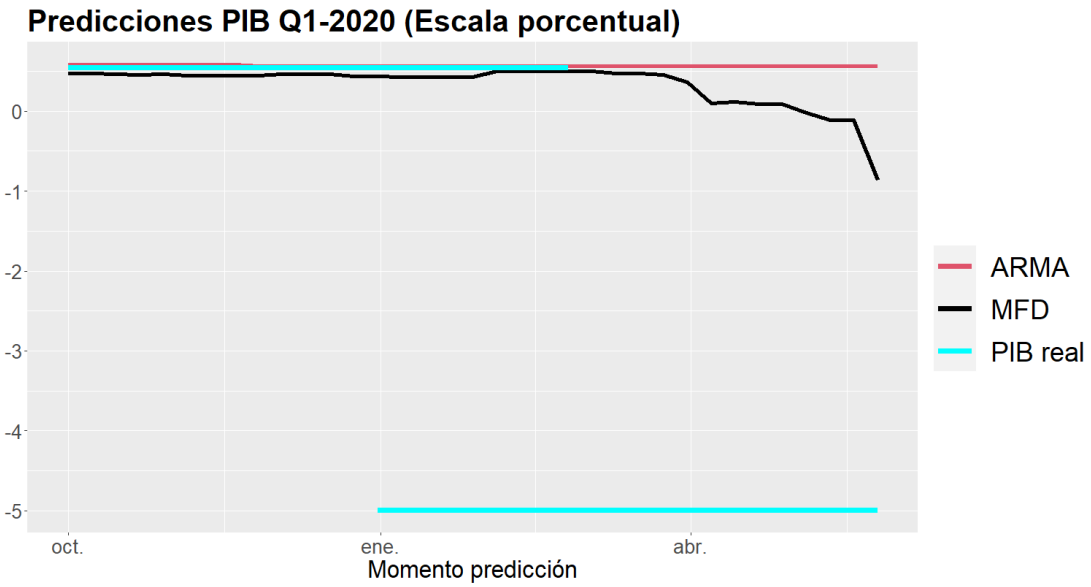

Figura 8.9: Predicción del PIB para el $Q1$ de 2020 obtenido a partir del panel tal como se tendría en los instantes indicados, teniendo en cuenta el retraso de publicación de las distintas variables. El intervalo donde se realizan las predicciones va de octubre de 2019 a mayo de 2020, realizando de esta forma *forecast* de octubre a diciembre, *nowcast* de enero a marzo y *backcast* de marzo a mayo. La línea cian representa el verdadero valor del PIB correspondiente a cada trimestre, la línea negra la predicción obtenida a partir del MFD y la línea roja la predicción obtenida a partir del modelo de referencia, el modelo ARMA, con su orden seleccionado en cada instante empleando el criterio BIC y parámetros estimados por máxima verosimilitud. Cada línea para el PIB se extiende desde su trimestre relativo hasta el instante en que se publica su valor teniendo el cuenta el retraso de publicación de 50 días.

8.6. Conclusiones finales

Tras examinar y desglosar los distintos parámetros, enfoques y técnicas en el marco del MFD, y una vez implementado en nuestro caso práctico para predecir el PIB de Galicia a partir de variables del entorno macroeconómico, hay varias conclusiones que se han ido destacando a lo largo del trabajo y que resulta interesante recopilar. Lo primero que cabe incluir en esta sección es la filosofía intrínseca al MFD. A partir de un conjunto de variables de gran tamaño (tanto temporalmente como en el número de series consideradas), se condensa la información más relevante sobre todas ellas y se resume a partir de una serie de variables latentes, los factores. De esta forma se tiene una reducción de la dimensión, pudiendo emplear los factores para interpretar el panel global de variables o en regresiones sucesivas para efectuar predicciones de cualquier variable de interés. Además de la simplificación que ofrece el MFD ante problemas de alta dimensión, otra propiedad importante es que es aplicable sobre paneles donde el resto de modelos empleados en economía no se pueden considerar debido a los problemas computacionales o de convergencia originados por el tamaño del panel.

En economía en particular se trabaja con cientos de variables de lo más diversas, por lo que la justificación para este modelo es clara en este campo. Otra propiedad importante es la sencillez del modelo, algo que interesa en economía. Una vez estimados los factores se dota al investigador de la información más relevante del entorno económico en un pequeño número de variables, lo que facilita la interpretación del contexto económico gracias a la reducción en la dimensión. Con los métodos clásicos, estamos obligados a no considerar gran parte de la información disponible, al poder trabajar solo con una o pocas series al unísono. Esto de nuevo es difícilmente justificable en econometría, un campo donde se cuenta con numerosas series las cuales suelen estar relacionadas de alguna forma. Debido a todo esto, son muchos los autores que han desarrollado el MFD durante los últimos años y muchas las entidades que lo comienzan a implementar.

En esta memoria se profundiza en la comparación entre el MFD y los métodos clásicos, comparando su comportamiento con el método de referencia, el ARMA. En la Sección 8.5 se realiza una comparación en donde el MFD obtiene resultados claramente mejores y deja patente toda la filosofía detrás de su construcción. Esta comparación se realiza en uno de los escenarios más adversos de cara a realizar estimaciones dado lo atípico y repentino cambio en el contexto económico. El MFD, al considerar toda la información disponible en cada momento va aproximando sus predicciones al valor verdadero conforme avanza el tiempo y se obtienen más datos. A pesar de que la predicción no es todo lo ajustada que debería ser, mejora claramente el modelo de referencia. Además, dicha mejora, se produce en todos los instantes considerados. Como ya se comentó, la Sección 8.5 puede que sea la más importante al resumir toda la filosofía del MFD en una gráfica y comprobar su buen funcionamiento en la práctica.

En los últimos años está habiendo un aumento en el desarrollo y utilización de estos modelos, por lo que aún tienen campo de mejora. Uno de los puntos a mejorar y por el que probablemente no se emplean más es por el escaso desarrollo que tienen en los software estadísticos, dificultando su implementación. Concluimos que el MFD es un modelo que se adapta perfectamente al ámbito económico y que resulta de utilidad tanto para la interpretación del contexto económico como para efectuar predicciones, dos de los principales objetivos en este campo. Además de la propia justificación teórica, la cual motiva a emplear el MFD en econometría, se comprueba el mejor comportamiento respecto a metodologías clásicas como la Box-Jenkins, dejando patente la mejora práctica que también se obtiene.

Bibliografía

[Angelini *et al.* (2008)] Angelini, E., Banbura, M., y Rünstler, G. (2008). Estimating and forecasting the euro area monthly national accounts from a dynamic factor model. *ECB Working Paper, nº 953.*

[Arouba *et al.* (2007)] Aruoba, S. B., Diebold, F. X., y Scotti, C. (2007). Real-time measurement of business conditions. *FRB International Finance Discussion Paper*, 901, 07-028.

[Bai (2003)] Bai, J. (2003). Inferential theory for factor models of large dimensions. *Econometrica*, 71(1), 135-171.

[Bai y Ng (2002)] Bai, J., y Ng, S. (2002). Determining the number of factors in approximate factor models. *Econometrica*, 70(1), 191-221.

[Bai y Ng (2007)] 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 y Ng (2008)] Bai, J., y Ng, S. (2008). Foundations and Trends in Econometrics. *Foundations and Trends in Econometrics*, 3(2), 89-163.

[Barhoumi *et al.* (2018)] Barhoumi, K., Darne, O. y Ferrara, L. (2014) Dynamic factor models: a review of the literature. *Journal of Business Cycle Measurement and Analysis*, 8(2).

[Bernanke *et al.* (2005)] 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.

[Boivin y Ng (2006)] Boivin, J., y Ng, S. (2006). Are more data always better for factor analysis?. *Journal of Econometrics*, 132(1), 169-194.

[Camacho y Doménech (2012)] Camacho, M., y Doménech, R. (2012). MICA-BBVA: a factor model of economic and financial indicators for short-term GDP forecasting. *SERIEs*, 3(4), 475-497.

[Camacho y Pérez-Quiros (2009)] Camacho, M., y Perez-Quiros, G. (2009). Ñ-Sting: España short term indicator of growth. *Cuadernos de Trabajo, Banco de España,* Nº 0912.

[Camacho y Doménech (2012)] Camacho, M., y Doménech, R. (2012). MICA-BBVA: a factor model of economic and financial indicators for short-term GDP forecasting. *SERIEs*, 3(4), 475-497.

[Cuevas *et al.* (2017)] Cuevas, A., Pérez-Quirós, G., y Quilis, E. M. (2017). Integrated model of short-term forecasting of the Spanish economy (MIPred model). *Revista de Economía Aplicada*, 25(74), 5-25.

[Cuevas y Quilis (2012)] Cuevas, Á., y Quilis, E. M. (2012). A factor analysis for the Spanish economy. *SERIEs*, 3(3), 311-338.

[de Valk *et al.* (2019)] de Valk, S., de Mattos, D., y Ferreira, P. (2019). Nowcasting: An R Package for Predicting Economic Variables Using Dynamic Factor Models. *The R Journal*, 11(1), 230-244.

[de Mattos *et al.* (2019)] Daiane Marcolino de Mattos, Pedro Costa Ferreira, Serge de Valk and Guilherme Branco Gomes (2019). nowcasting: Predicting Economic Variables using Dynamic Factor Models. R package version 1.1.4. https://CRAN.R-project.org/package=nowcasting

[Doz *et al.* (2011)] 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.

[Forni *et al.* (2005)] Forni, M., Hallin, M., Lippi, M., y Reichlin, L. (2005). The generalized dynamic factor model: one-sided estimation and forecasting. *Journal of the American Statistical Association*, 100(471), 830-840.

[Forni *et al.* (2007)] Forni, M., Giannone, D., Lippi, M.,y Reichlin, L. (2009). Opening the black box: Structural factor models with large cross sections. *Econometric Theory*, 1319-1347.

[Fuller (1996)] Fuller, W. A. (2009). *Introduction to Statistical Time Series* (Vol. 428). John Wiley & Sons.

[Giannone *et al.* (2008)] 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.

[Giannone y Reichlin (2006)] Doz, C., Giannone, D., y Reichlin, L. (2006). *A Quasi Maximum Likelihood Approach for Large Approximate Dynamic Factor Models. European Central Bank.* Working Paper. 674.(Septiembre).

[Holmes *et al.* (2012)] Holmes, E. E., Ward, E. J., y Wills, K. (2012). MARSS: Multivariate Autoregressive State-space Models for Analyzing Time-series Data. *R journal*, 4(1).

[Holmes *et al.* (2014)] Holmes, E. E., Ward, E. J., y Scheuerell, M. D. (2014). Analysis of multivariate time-series using the MARSS package. *NOAA Fisheries, Northwest Fisheries Science Center*, 2725, 98112.

[Holmes *et al.* (2020)] Elizabeth Holmes, Eric Ward, and Kellie Wills (2020). MARSS: Multivariate Autoregressive State-Space Modeling. R package version 3.10.12.

[Johansen (1988)] Johansen, S. (1988). Statistical analysis of cointegration vectors. *Journal of economic dynamics and control*, 12(2-3), 231-254.

[Mariano y Murasawa (2003)] 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.

[Melo *et. al* (2005)] Melo-Velandia, L. F., Ramírez-Cortés, J. M., Ramos-Veloza, M. A., Melo-Velandia, L. F., y Ramos-Veloza, M. A. (2005). Construcción de un "índice de percepción de riesgo" de los mercados financieros globales. *Borradores de Economía*; No. 344.

[Pateiro B. (2019)] Análisis Multivariante. *Apuntes de la asignatura, Universidad de Santiago de Compostela*

[R Core Team (2018)] R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

[Reis y Watson (2010)] Reis, R., y Watson, M. W. (2010). Relative goods' prices, pure inflation, and the Phillips correlation. *American Economic Journal: Macroeconomics*, 2(3), 128-57.

[Sargent y Sims (1977)] Sargent, T. J., y Sims, C. A. (1977). Business cycle modeling without pretending to have too much a priori economic theory. *New methods in business cycle research*, 1, 145-168.

[Sosa-Escudero (1997)] Sosa Escudero, W. (1997). Testing for unit-roots and trend-breaks in Argentine real GDP. *Económica*, 43.

[Stock y Watson (1998a)] Stock, J. H., y Watson, M. W. (1998). Diffusion indexes. *NBER working paper*, (w6702).

[Stock y Watson (1998)] Stock, J. H., y Watson, M. W. (1988). A probability model of the coincident economic indicators (No. w2772). *National Bureau of Economic Research*.

[Stock y Watson (2002)] Stock, J. H., y Watson, M. W. (2002). Macroeconomic forecasting using diffusion indexes. *Journal of Business & Economic Statistics*, 20(2), 147-162.

[Stock y Watson (2002a)] Stock, J. H., y Watson, M. W. (2002). Forecasting using principal components from a large number of predictors. *Journal of the American Statistical Association*, 97(460), 1167-1179.

[Stock y Watson (2005)] Stock, J. H., y Watson, M. W. (2005). Implications of dynamic factor models for VAR analysis (No. w11467). *National Bureau of Economic Research*.

[Stock y Watson (2006)] Stock, J. H., y Watson, M. W. (2006). Forecasting with many predictors. *Handbook of Economic Forecasting*, 1, 515-554.

[Stock y Watson (2011)] Stock, J. H., y Watson, M. (2011). Dynamic factor models. *Oxford Handbooks Online*.

[Stock y Watson (2016)] Stock, J. H., y Watson, M. W. (2016). Dynamic factor models, factor-augmented vector autoregressions, and structural vector autoregressions in macroeconomics. *In Handbook of Macroeconomics* (Vol. 2, pp. 415-525). Elsevier.