

Proyecto Series de Tiempo: Dynamic Factor Models

Sacbe García García, David Alejandro Muro Campa

Centro de Investigación en Matemáticas, A.C.
Unidad Monterrey
Monterrey, Nuevo León
Series de Tiempo.

9 de diciembre de 2023

INTRODUCCIÓN

Motivación

Los modelos factoriales ofrecen una perspectiva valiosa en el análisis de variables económicas, como se evidencia al descomponer el comportamiento de una variable económica (x_{it}) en componentes específicos. Estos modelos, impulsados por factores inobservables (f_t), comunes a todas las variables pero con efectos específicos (λ_i), capturan la complejidad subyacente en los datos.

Desde una perspectiva económica, la noción de fuerzas comunes que influyen en todas las variables resulta atractiva, especialmente en contextos como el ciclo económico real y el equilibrio genérico dinámico estocástico. En estas instancias, unas pocas perturbaciones clave afectan globalmente a variables como la productividad, la demanda y la oferta, mientras se permiten perturbaciones adicionales específicas para cada variable.

La capacidad de los modelos factoriales para manejar conjuntos de datos extensos (N grandes) se alinea con la creciente disponibilidad de información masiva utilizada por responsables políticos y agentes económicos. Esta característica resuelve la maldición de la dimensionalidad inherente a los VAR estándar, donde el número de parámetros crece cuadráticamente con el número de variables. Además, estos modelos mitigan el sesgo de las variables omitidas y abordan problemas de no fundamentalidad en las perturbaciones, evitando dependencias de información futura no recuperables adecuadamente en los VAR. Su robustez frente a rupturas estructurales añade un nivel adicional de confiabilidad, mientras que las condiciones mínimas para los errores ofrecen flexibilidad al tolerar correlaciones temporales y heteroscedasticidad.

Contexto Histórico

El análisis de factores es una técnica de reducción de dimensión que resume las fuentes de variación entre las variables. El método se introdujo en 1904 por Spearman [1], para describir la inteligencia de los individuos con una variable no observada que se denominó factor. El método fue extendido por Geweke en 1977 [2] para capturar los co-movimientos en series de tiempo económicas.

Este se basa en un concepto muy importante en macroeconomía es que varios indicadores económicos tienden a moverse de manera sistemática a lo largo del tiempo, reflejando las diferentes fases del ciclo económico. Diferentes variables macroeconómicas, como el producto interno bruto (PIB), el empleo, la inversión y el gasto del consumidor, a menudo muestran un alto grado de correlación. Durante las expansiones económicas, estas variables tienden a aumentar juntas, mientras que durante las contracciones o recesiones tienden a disminuir juntas.

Las primeras aplicaciones datan en los 70s, en esos años los modelos VAR y VARMA sufrían de la maldición de la dimensio-

nalidad. Un modelo simple VAR(1) n -dimensional requiere no menos de $n(3n+1)/2$ parámetros (n^2 coeficientes autorregresivos, n varianzas de las innovaciones y $n(n-1)/2$ covarianzas de las innovaciones). Geweke [2], Sargent y Sims [3] fueron los primeros en 1977 en introducir el modelo y los supuestos que son más débiles que el modelo de factores clásico. La idea más importante de Geweke fue la idea de lo que conocemos como *cargas dinámicas*. Esta idea junto con los supuestos conforma el modelo de factores dinámicos exacto.

En 1903 Chamberlain [4] y Chamberlain y Rothschild [5] aprovechan la característica de la alta dimensionalidad de los datos en la práctica, introduciendo los modelos de factores dinámicos aproximados.

Estas son las bases de la primera explosión que experimentaría el campo en los primeros años de los 2000 debido principalmente 3 grupos de econométristas, Stock y Watson en 2002 [7] [8], Bai y Serena Ng en 2002 y 2003 [6], [9] y Forni, Hallin, Lippi y Reichlin en el 2000 [10]. En este proyecto revisamos los enfoques principalmente de los primeros dos grupos, el último grupo introduce el *General Dynamic Factor Model* que no se discutirá aquí.

Estas contribuciones han disparado los artículos en el tema, refinando y extendiendo las aplicaciones más allá de la econometría y las finanzas como lo son las ciencias ambientales [12], medicina [13] y de nuevo a los orígenes en psicología [14]. Para un contexto histórico más a fondo se puede consultar [11]. La Figura (1) muestra un "árbol genealógico" de el desarrollo de los modelos de factores dinámicos.

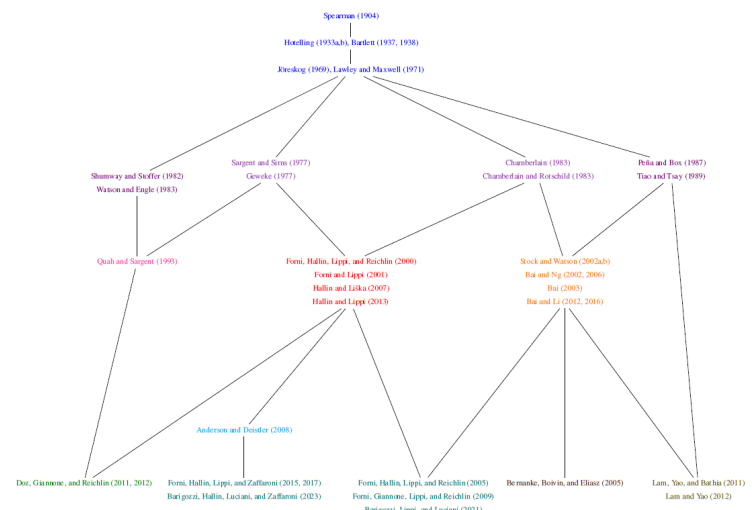


Figura 1. Genealogía de los Dynamic Factor Models [11].

MODELOS CON FACTORES

Modelo de Factores Exacto

El modelo de factores exacto clásico fue introducido por Spearman [1]. El modelo asume que una observación x_{it} se puede descomponer en la suma de dos componentes mutuamente ortogonales inobservadas: la componente común $\beta_{it} = \lambda_i' f_t$ y una componente idiosincrática $\zeta = \mu_i + e_{it}$ donde $i = 1, \dots, n$ con n el número de variables, $t = 1, \dots, T$ donde T es el número de observaciones disponibles de cada variable.

Las componentes idiosincráticas se asumen que no están correlacionadas de tal forma que ζ_{ks} y ζ_{it} son mutuamente ortogonales para todo $k \neq i$ y cualquier s y t , por lo tanto, toda la correlación entre las variables observables es debida a los factores. El modelo se puede escribir como

$$x_{it} = \mu_i + \sum_{j=1}^r \lambda_{ij} f_{jt} + e_{it}, \quad i = 1, \dots, n, t = 1, \dots, T \quad (1)$$

donde f_{jt} y e_{it} son ruidos blancos ortogonales para cualquier i y j ; e_{it} y e_{kt} son ortogonales para $k \neq i$ y λ_{ij} son las cargas del j -ésimo factor en la i -ésima variable. La Ecuación (1) se puede escribir como

$$x_{it} = \mu_i + \lambda_i' f_t + e_{it}, \quad i = 1, \dots, n, t = 1, \dots, T \quad (2)$$

con $\lambda_i' = (\lambda_{i1}, \dots, \lambda_{ir})$ y $f_t = (f_{1t}, \dots, f_{rt})'$. El componente común es $\beta_{it} = \sum_{j=1}^r \lambda_{ij} f_{jt}$, y la componente idiosincrática es $\zeta_{it} = \mu_i + e_{it}$.

En notación matricial con $x_t = (x_{1t}, \dots, x_{nt})'$, $f_t = (f_{1t}, \dots, f_{rt})'$ y $e_t = (e_{1t}, \dots, e_{nt})'$, el modelo se puede escribir como

$$x_t = \mu + \Lambda f_t + e_t \quad (3)$$

donde $\Lambda = (\lambda_1, \dots, \lambda_n)'$ es una matriz $n \times r$ de rango completo. La matriz de covarianzas de e_t es diagonal dado que los términos idiosincráticos se asumen no están correlacionados.

Mientras el supuesto principal es que los procesos f_t y e_t son ortogonales entre sí, el modelo exacto de factores asume que los componentes idiosincráticos son también ortogonales entre sí, de tal manera que la correlación entre las variables observables es puramente debida a los factores comunes. Ambos supuestos son necesarios para que el modelo sea identificable.

Los modelos de factores se introdujeron en economía por Geweke [2] y Sargent y Sims [3]. Estos autores generalizaron el modelo para capturar la dinámica en los datos. Usando la misma notación que en la Ecuación (3), el modelo exacto dinámico de factores se puede escribir como

$$x_t = \mu + \Lambda_0 f_t + \Lambda_1 f_{t-1} + \dots + \Lambda_s f_{t-s} + e_t \quad (4)$$

o de una forma más compacta

$$x_t = \mu + \Lambda(L) f_t + e_t \quad (5)$$

donde L es el operador de retroceso.

El modelo admite una representación estática

$$x_t = \mu + \Lambda F_t + e_t \quad (6)$$

donde $F_t = (f_t', f_{t-1}', \dots, f_{t-s}')'$.

Las representación (5) captura la dependencia de las variables observadas en los lags de los factores explícitamente, mientras que la representación estática (6) captura la dependencia implícitamente.

Modelo de Factores Aproximado

Como ya se ha señalado, los modelos factoriales exactos se basan en un supuesto muy estricto de ausencia de correlación cruzada entre los componentes idiosincráticos. En dos artículos fundamentales, Chamberlain [4] y Chamberlain y Rothschild [5] introdujeron modelos factoriales aproximados relajando este supuesto. Permitieron que los componentes idiosincráticos estuvieran ligeramente correlacionados y proporcionaron un conjunto de condiciones que garantizaban que los modelos factoriales aproximados se identificaran asintóticamente como $n \rightarrow \infty$.

Sea $x_t^n = (x_{1t}, \dots, x_{nt})'$ el vector que contiene la t^{th} observación de las primeras n variables como $n \rightarrow \infty$, y $\Sigma_n = \text{cov}(x_t^n)$ es la matriz de covarianza de x_t . Denotando por $\lambda_1(\mathbf{A}) \geq \lambda_2(\mathbf{A}) \geq \dots \geq \lambda_n(\mathbf{A})$ los valores propios ordenados de cualquier matriz simétrica \mathbf{A} con tamaño $(n \times n)$, los supuestos subyacentes a los modelos factoriales aproximados son los siguientes:

- (CR1) $\text{Sup}_n \lambda_r(\Sigma_n) = \infty$ (los r mayores valores propios de Σ_n son divergentes)
- (CR2) $\text{Sup}_n \lambda_{r+1}(\Sigma_n) < \infty$ (los restantes valores propios de Σ_n están acotados)
- (CR3) $\text{Inf}_n \lambda_n(\Sigma_n) > 0$ (Σ_n no se aproxima a la singularidad)

Los autores demuestran que bajo los supuestos (CR1)-(CR3), existe una descomposición única $\Sigma_n = \Lambda_n \Lambda_n' + \Psi_n$, donde Λ_n es una secuencia de matrices anidadas $n \times r$ con $\text{rank } r$ y $\lambda_i(\Lambda_n \Lambda_n') \rightarrow \infty, \forall i = 1 \dots r$, y $\lambda_1(\Psi_n) < \infty$. Alternativamente, x_t puede descomponerse utilizando un par de procesos vectoriales aleatorios f_t y e_t^n mutuamente ortogonales

$$x_t^n = \mu_n + \Lambda_n f_t + e_t^n$$

con $\text{cov}(f_t) = \mathbf{I}_r$ y $\text{cov}(e_t^n) = \Psi_n$, donde los r factores comunes son omnipresentes, en el sentido de que el número de variables afectadas por cada factor crece con n , y los términos idiosincráticos pueden estar ligeramente correlacionados con covarianzas acotadas.

Al igual que los modelos de factores dinámicos exactos, los modelos de factores dinámicos aproximados también se basan en una ecuación que vincula las series observables con los factores y sus retardos, pero en este caso los términos idiosincráticos pueden estar ligeramente correlacionados entre sí y se supone que el número de series tiende a infinito. El modelo tiene una dinámica

$$x_t^n = \mu_n + \Lambda_n(L) f_t + e_t^n$$

y una representación estática

$$x_t^n = \mu_n + \Lambda_n F_t + e_t^n$$

equivalentes a 5, y 6 respectivamente.

ESTIMACIÓN

Análisis de Componentes Principales

Chamberlain y Rothschild [5] sugirieron utilizar el análisis de componentes principales (PCA) para estimar el modelo factorial estático aproximado, y Stock y Watson [8] y Bai y Ng [9] popularizaron este enfoque en macroeconometría.

Considerando datos centrados y asumiendo que el número de factores, r , es conocido, PCA permite estimar simultáneamente los factores y sus cargas resolviendo el problema de mínimos cuadrados

$$\begin{aligned} \min_{\Lambda, \tilde{\mathbf{F}}} \frac{1}{nT} \sum_{i=1}^n \sum_{t=1}^T (x_{it} - \lambda'_i \tilde{\mathbf{f}}_t)^2 = \\ \min_{\Lambda, \tilde{\mathbf{F}}} \frac{1}{nT} \sum_{t=1}^T (\mathbf{x}_t - \Lambda \tilde{\mathbf{f}}_t)' (\mathbf{x}_t - \Lambda \tilde{\mathbf{f}}_t). \end{aligned} \quad (7)$$

Debido a la mencionada indeterminación rotacional de los factores y sus cargas, las estimaciones de los parámetros deben ser restringidas para obtener una solución única. Generalmente, se impone una de las dos condiciones de normalización siguientes

$$\frac{1}{T} \sum_{t=1}^T \hat{\mathbf{f}}_t \hat{\mathbf{f}}_t' = \mathbf{I}_r \quad \text{o} \quad \frac{\hat{\Lambda}' \hat{\Lambda}}{n} = \mathbf{I}_r$$

Utilizando la primera normalización y eliminando Λ , se obtiene una matriz de factores estimada, $\hat{\mathbf{F}}$, que es T veces los eigenvectores correspondientes a los r mayores eigenvalores de la matriz $T \times T$ $\mathbf{X}\mathbf{X}'$. Dada $\hat{\mathbf{F}}$, $\hat{\Lambda} = (\hat{\mathbf{F}}' \hat{\mathbf{F}})^{-1} \hat{\mathbf{F}}' \mathbf{X} = \hat{\mathbf{F}}' \mathbf{X} / T$ es la matriz correspondiente de cargas de factores. La solución al problema de minimización anterior no es única, aunque la suma de residuos al cuadrado lo es. Otra solución esta dada por $\tilde{\Lambda}$ construida como n veces los eigenvectores correspondientes a los r mayores eigenvalores de la matriz $n \times n$ $\mathbf{X}'\mathbf{X}$. Utilizar la segunda normalización aquí implica que $\tilde{\mathbf{F}} = \mathbf{X}\tilde{\Lambda}/n$. Bai y Ng [9] indicaron que el último enfoque es computacionalmente menos costoso cuando $T > n$, mientras que el primero es menos exigente cuando $T < n$. En ambos casos, los componentes idiosincráticos se estiman por $\tilde{\mathbf{e}}_t = \mathbf{x}_t - \hat{\Lambda} \hat{\mathbf{f}}_t$, y su covarianza se estima mediante la matriz de covarianza empírica de $\tilde{\mathbf{e}}_t$.

Análisis de Componentes Principales Generalizados

La estimación generalizada de componentes principales imita a mínimos cuadrados generalizados para tratar con una matriz de varianza no esférica de los componentes idiosincráticos. Aunque el método se había utilizado anteriormente, Choi [15] fue quien demostró que se pueden lograr ganancias de eficiencia al incluir una matriz de ponderación en la minimización

$$\min_{\Lambda, \tilde{\mathbf{F}}} \frac{1}{nT} \sum_{t=1}^T (\mathbf{x}_t - \Lambda \tilde{\mathbf{f}}_t)' \Psi^{-1} (\mathbf{x}_t - \Lambda \tilde{\mathbf{f}}_t)$$

El estimador factible generalizado de componentes principales reemplaza la Ψ desconocida por un estimador $\tilde{\Psi}$. Los elementos diagonales de Ψ pueden estimarse mediante las varianzas de los términos idiosincráticos individuales. Bai y Liao [17] derivan las condiciones bajo las cuales el estimador $\tilde{\Psi}$ puede tratarse como conocido. El problema de minimización ponderado anterior se basa en la suposición de choques idiosincráticos independientes, lo cual puede ser demasiado restrictivo en

la práctica. Stock y Watson [16] aplicaron un filtro diagonal $\mathbf{D}(L)$ a los términos idiosincráticos para tratar la correlación serial, de modo que el problema se convierte en

$$\min_{\mathbf{D}(L), \Lambda, \tilde{\mathbf{F}}} \frac{1}{T} \sum_{t=1}^T (\mathbf{D}(L)\mathbf{x}_t - \Lambda \tilde{\mathbf{f}}_t)' \tilde{\Psi}^{-1} (\mathbf{D}(L)\mathbf{x}_t - \Lambda \tilde{\mathbf{f}}_t),$$

donde $\tilde{\mathbf{F}} = (\tilde{\mathbf{f}}_1 \dots \tilde{\mathbf{f}}_T)'$ con $\tilde{\mathbf{f}}_t = \mathbf{D}(L)\mathbf{f}_t$ y $\tilde{\Psi} = E[\tilde{\mathbf{e}}_t \tilde{\mathbf{e}}_t']$ con $\tilde{\mathbf{e}}_t = \mathbf{D}(L)\mathbf{e}_t$. La estimación de $\mathbf{D}(L)$ y $\tilde{\Psi}$ puede hacerse de manera secuencial, iterando hasta la convergencia. Breitung y Tenhofen (2011) proponen un procedimiento de estimación de dos pasos similar que permite términos idiosincráticos heterocedásticos y correlacionados serialmente.

Método Two-step

Doz et.al. [25] propusieron un estimador a dos pasos que toma en cuenta la dinámica de los factores. Asumieron que los factores en el modelo de factores aproximado sigue un proceso VAR (vector autoregressive process):

$$\mathbf{x}_t = \Lambda \mathbf{f}_t + \mathbf{e}_t \quad (8)$$

$$\mathbf{f}_t = \Phi_1 \mathbf{f}_{t-1} + \dots + \Phi_p \mathbf{f}_{t-p} + \mathbf{u}_t \quad (9)$$

Permiten que los términos idiosincráticos sean autocorrelacionados pero no especifican su dinámica. El proceso de estimación es el siguiente:

1. Estimadores preliminares de las cargas $\hat{\Lambda}$ y de los factores, $\hat{\mathbf{f}}_t$ son obtenidos mediante análisis de componentes principales. Los términos idiosincráticos se estiman como $\hat{\mathbf{e}}_{it} = x_{it} - \hat{\lambda}'_i \hat{\mathbf{f}}_t$ y su varianza es estimada mediante la varianza empírica asociada $\hat{\psi}_{ii}$. Los factores estimados, $\hat{\mathbf{f}}_t$ son usados en un modelo VAR para obtener las estimaciones $\hat{\Phi}_j, j = 1, \dots, p$
2. El modelo se expresa en la forma espacio-estado de la siguiente forma, con $s < p - 1$

$$\begin{bmatrix} \mathbf{f}_1 \\ \mathbf{f}_{t-1} \\ \vdots \\ \mathbf{f}_{t-p+1} \end{bmatrix} = \begin{bmatrix} \Phi_1 & \Phi_2 & \dots & \Phi_p \\ \mathbf{I}_q & \mathbf{0} & \dots & \mathbf{0} \\ \vdots & \ddots & \ddots & \vdots \\ \mathbf{0} & \dots & \mathbf{I}_q & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{f}_{t-1} \\ \mathbf{f}_{t-2} \\ \vdots \\ \mathbf{f}_{t-p} \end{bmatrix} + \begin{bmatrix} \mathbf{I}_q \\ \mathbf{0} \\ \vdots \\ \mathbf{0} \end{bmatrix} \mathbf{u}_t \quad (10)$$

y $\text{cov}(\mathbf{u}_t) = \mathbf{I}_r$ y $\text{cov}(\mathbf{e}_t) = \Psi = \text{diag}(\psi_{11}, \dots, \psi_{nn})$. Usando las estimaciones de los parámetros $\hat{\Lambda}, \hat{\Psi}, \hat{\Phi}_j, j = 1, \dots, p$ obtenidos en el primer paso, una ejecución del suavizador de Kalman es aplicado a los datos. Esto produce un nuevo estimador del factor $\mathbf{f}_{t|T} = E(\mathbf{f}_t | \mathbf{x}_1, \dots, \mathbf{x}_T, \hat{\theta})$, donde $\hat{\theta}$ es un vector que contiene las estimaciones del primera paso de todos los parámetros.

Doz et al. [25] demostraron que bajo sus supuestos, los parámetros estimados $\hat{\theta}$ son consistentes y convergen a un $\text{rate} (\min(\sqrt{n}, \sqrt{T}))^{-1}$

DETERMINAR NÚMERO DE FACTORES

Criterio de Bai y Ng

Para especificar completamente un modelo de factores dinámicos se requiere seleccionar el número de factores comunes. El número de factores estáticos puede ser determinado por un criterio de información que puede construirse mediante funciones objetivo o analizando la distribución de eigenvalores. Bai y Ng [9] usaron un criterio de información de la forma $IC(r) = \ln V_r(\hat{\Lambda}, \hat{F}) + rg(n, T)$, donde $V_r(\hat{\Lambda}, \hat{F})$ es la función objetivo de mínimos cuadrados (7) evaluada con los estimadores de componentes principales cuando r factores son considerados, y $g(n, T)$ es una función de penalización que satisface dos condiciones: $g(n, T) \rightarrow 0$ y $\min[n^{1/2}, T^{1/2}] \cdot g(n, T) \rightarrow \infty$ conforme $n, T \rightarrow \infty$.

El estimador del número de factores es $\hat{r}_{IC} = \min_{0 \leq r \leq r_{max}} IC(r)$ donde r_{max} es el límite superior del valor verdadero de factores. Los autores mostraron que el estimador es consistente sin restricciones entre n y T y el resultado es robusto para cuando hay heterocedasticidad y correlación cruzada.

Criterio de Ahn-Horenstein-Onatski

Ahn y Horenstein [18] y Onatski [19, 20] adoptan un enfoque diferente al comparar los eigenvalores adyacentes de la matriz de densidad espectral en una frecuencia dada o de la matriz de covarianza de los datos. La idea básica detrás de este enfoque es que los primeros r eigenvalores serán no acotados, mientras que los valores restantes serán acotados. Por lo tanto, la proporción de los eigenvalores subsiguientes se maximiza en la ubicación del mayor salto relativo en un gráfico de barras (un gráfico de los eigenvalores ordenados contra el rango de esos eigenvalores). Estos autores también presentan estadísticas alternativas utilizando la diferencia, la proporción de cambios y las tasas de crecimiento de los eigenvalores subsiguientes.

Onatski proporciona un límite conjunto (una ley vectorial de Tracy-Widom) para los primeros varios eigenvalores centrados y escalados. Dado que el problema es simétrico en N y T , este resultado se aplica igualmente a paneles con correlación transversal pero sin dependencia de series temporales. Onatski [19] desarrolla una estadística pivotal que se puede utilizar para probar la hipótesis de que $q = q_0$ contra la hipótesis de que $q > q_0$. El test de Onatski está en la línea del análisis de componentes principales generalizadas de Brillinger en el sentido de que examina una función de los eigenvalores de la matriz de densidad espectral suavizada en una frecuencia dada, donde la función se elige de manera que la estadística de prueba sea pivotal.

FORECASTING

Uno de los usos mas importantes de los modelos de factores dinámicos es predicción. Predicción usando modelos de factores de alta dimensión fueron introducidos en la literatura por Stock y Watson [7]. El método al que llamaron *diffusion index forecast* consiste en estimar los factores \mathbf{f}_t por componentes principales y usar esas estimaciones en un modelo de regresión por mínimos cuadrados

$$\mathbf{y}_{t+h} = \beta'_f \hat{\mathbf{f}}_t + \beta'_w \mathbf{w}_t + \epsilon_{t+h} \quad (11)$$

donde y_t es la variable de interés, $\hat{\mathbf{f}}_t$ es el vector de los factores estimados y \mathbf{w}_t es un vector de predictores observables. La predicción directa para el tiempo $T + h$ es entonces

$$\mathbf{y}_{T+h|T} = \hat{\beta}'_f \hat{\mathbf{f}}_T + \hat{\beta}'_w \mathbf{w}_T \quad (12)$$

Los autores demostraron que bajo los supuestos que usaron para asegurar la consistencia de los estimadores de componentes principales

$$\text{plim}_{n \rightarrow \infty} \left[(\hat{\beta}'_f \hat{\mathbf{f}}_T + \hat{\beta}'_w \mathbf{w}_T) - (\beta'_f \mathbf{f}_T + \beta'_w \mathbf{w}_T) \right] = 0 \quad (13)$$

entonces la predicción es asintóticamente equivalente a la que hubiera sido si los factores hubieran sido observados. Además, bajo supuestos más fuertes Bai y Ng [26] demostraron que el error de predicción es asintóticamente gaussiano con varianza conocida entonces podemos calcular intervalos de confianza.

Bajo el método de two-step que describimos arriba Doz et al. [25], los estimadores del modelo VAR para \mathbf{f}_t pueden ser usados recursivamente para predecir $\hat{\mathbf{f}}_{T+h|T}$ al tiempo T . Una predicción consistente en media cuadrada para el período $T + h$ se puede obtener mediante $\mathbf{y}_{T+h|T} = \hat{\Lambda} \hat{\mathbf{f}}_{T+h|T}$. Stock y Watson [8] apunan que el método *diffusion index* y el enfoque *two-step* son equivalentes.

FACTORES COMO VARIABLES INSTRUMENTALES

Kapetanios y Marcellino [21] y Bai y Ng [22] consideran el uso de factores estimados como variables instrumentales. La motivación para hacerlo radica en que los factores condensan la información en un gran número de series, de modo que se puede realizar un análisis de variables instrumentales o de momentos generales (GMM) utilizando menos instrumentos y potencialmente más fuertes que si se utilizara los datos originales. La propuesta de utilizar los componentes principales de un gran número de series como instrumentos se remonta a Kloeck y Mennes [23] y Amemiya [24], sin embargo, los tratamientos tempranos requerían exogeneidad estricta de los instrumentos. La necesidad de exogeneidad estricta puede debilitarse si hay una estructura de factores y si F_t es un instrumento válido.

El resultado principal de Kapetanios y Marcellino [21] y Bai y Ng [22] es que, si F_t constituye un instrumento fuerte y $\hat{\beta}^{TSLS}(F)$ es el estimador de mínimos cuadrados en dos etapas basado en los instrumentos F , entonces bajo las condiciones de Bai-Ng (en particular, $N^2/T \rightarrow \infty$), $\sqrt{T} [\hat{\beta}^{TSLS}(F) - \hat{\beta}^{TSLS}(\hat{F})] \xrightarrow{p} 0$, es decir, el estimador de componentes principales puede tratarse como datos para el propósito de regresión de variables instrumentales con instrumentos fuertes. Extensiones de este resultado incluyen la irrelevancia de la estimación de F incluso si F_t es un instrumento débil, asumiendo que hay una estructura de factores.

APLICACIONES

Euro Area GDP

La eurozona, denominada oficialmente zona del euro y llamada también zona euro, es el conjunto de Estados de la Unión (UE) que han eliminado sus monedas nacionales, adoptando el euro como moneda oficial (20 Estados). En la Figura (2) se observan los países que conforman la eurozona.

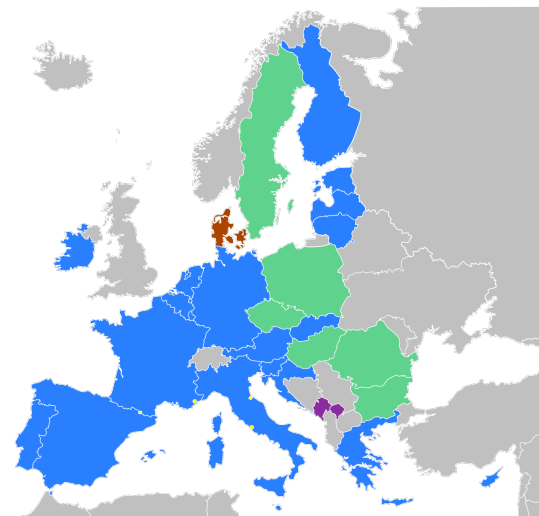


Figura 2. Países que conforman la zona del euro.

Se tiene un conjunto de datos de 92 indicadores macro-económicos como producción industrial, desempleo, PIB, entre otras desde enero de 1980 hasta septiembre de 2009. En la Figura (5) se aprecian las 92 series de tiempo.

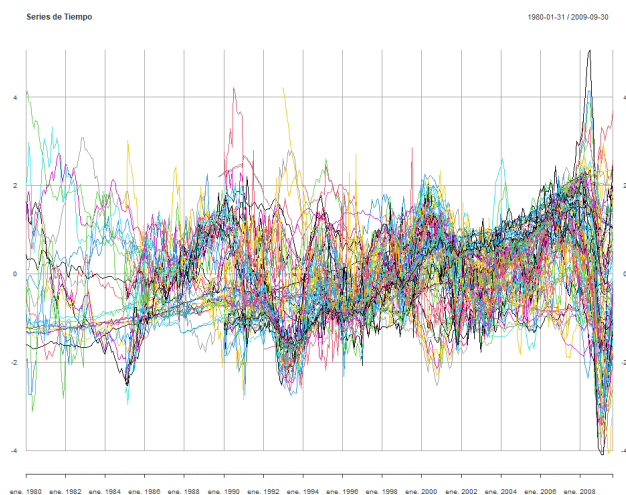


Figura 3. Series de Tiempo para cada una de las 92 variables indicadoras de la economía en los países de la eurozona.

Antes de ajustar el modelo debemos definir el número de factores a utilizar. Utilizando el criterio de Bai y Ng [9]

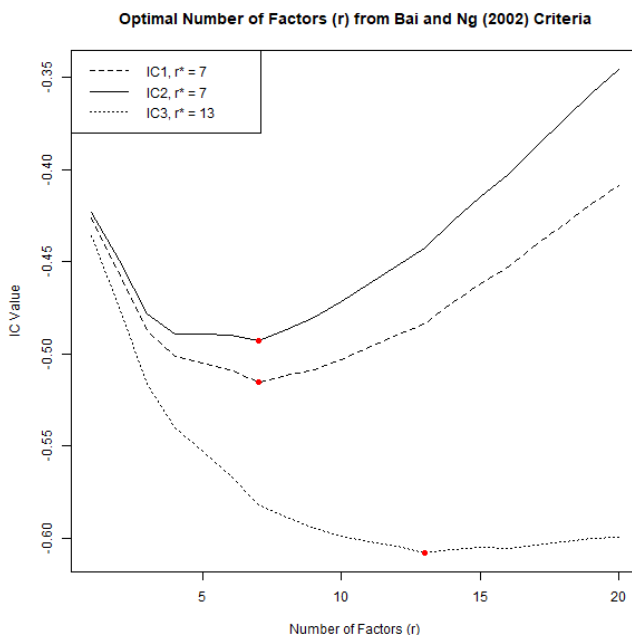


Figura 4. Número de Factores.

Por lo tanto elegimos 7 como número de factores. Para establecer el orden del proceso VAR(p) que seguirán los factores, utilizamos varias medidas como AIC, BIC, HQ, concluyendo que el orden debe ser $p = 3$. Ajustamos y obtenemos los factores que se muestran en la Figura (5)

Por último haciendo forecasting obtenemos los resultados que se muestran en la Figura (6)

Impacto de los datos FRED-QD en el GDPC1

El conjunto de datos FRED-QD se refiere a la Base de Datos de Datos Trimestrales (Quarterly Data) del Sistema de Reserva Federal de St. Louis (FRED). Este conjunto de datos proporciona información económica a nivel trimestral y abarca diversas variables que son relevantes para el análisis macroeconómico. Por otra parte el Real Gross Domestic Product (GDPC1) se refiere al Producto Interno Bruto (PIB) real de Estados Unidos. El PIB es una medida de la producción total de bienes y servicios dentro de los límites geográficos de un país durante un período de tiempo específico. La palabra real indica que el PIB se ajusta por cambios en el nivel general de precios, lo que significa que se tiene en cuenta la inflación o deflación. Este ejercicio utilizará el conjunto de datos de febrero de 2020, y luego se examinará cómo los datos entrantes actualizados en los meses de marzo a junio influyeron en el pronóstico del modelo para el crecimiento del PIB real en el segundo trimestre de 2020.

Como es típico en estos ejercicios, se transforman las variables para inducir estacionariedad. En particular, los conjuntos de datos FRED-MD y FRED-QD incluyen transformaciones sugeridas que suelen aplicar diferencias o transformaciones de cambio porcentual.

Siguiendo a McCracken y Ng, se eliminan los valores atípicos, definidos como observaciones que están más de 10 veces el rango intercuartílico desde la media de la serie. Para ver cómo funcionan la transformación y la eliminación de atípicos,

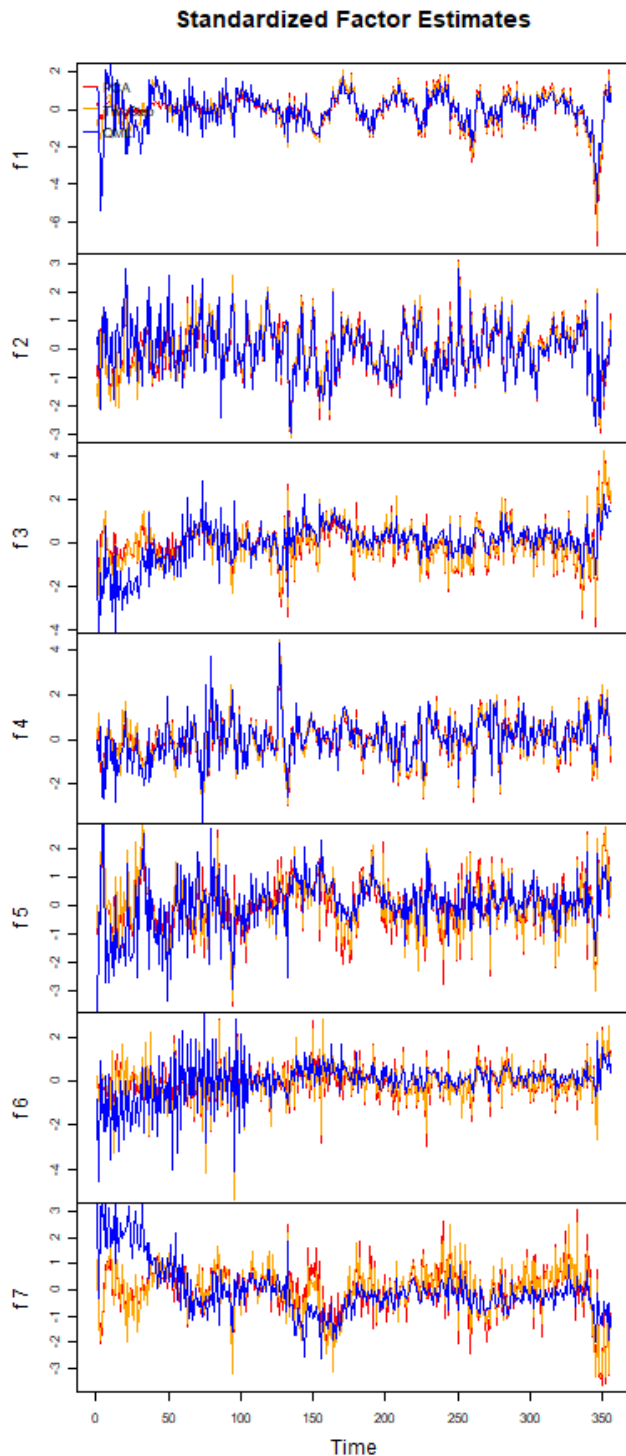


Figura 5. Factores estimados mediante PCA, Two-step y QML.

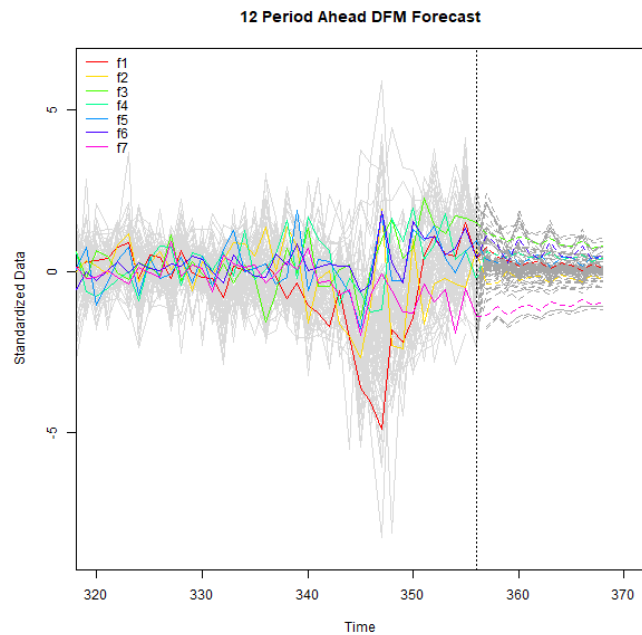


Figura 6. Forecasting.

se presentan tres gráficos de la variable RPI (Íngreso Personal Real”) durante el período de enero de 2000 a enero de 2020 como se muestra en la figura 7.

A continuación, se obtienen los grupos para cada serie a partir de los archivos de definición de FRED-MD y FRED-QD y luego se muestran cuántas de las series que se utilizan pertenecen a cada uno de los grupos como se muestra en la tabla 1.

Grupo	Número de Series
Output and Income	17
Consumption, orders, and inventories	10
Labor Market	31
Housing	10
Money and credit	13
Stock market	5
Interest and exchange rates	22
Prices	20

Tabla 1. Número de series en cada grupo.

Ahora se necesita especificar los detalles del modelo dinámico de factores específico que se quiere estimar. En particular: el número de factores a incluir, qué variables cargan en qué factores y el orden de la autoregresión según la cual evoluciona el factor; si estandarizar o no el conjunto de datos para la estimación y si modelar los términos de error idiosincráticos como procesos AR(1) o como iid. Si bien hay varias formas de identificar cuidadosamente el número de factores que se deben usar como los mostrados anteriormente según Fulton [27] una estructura adecuada es dos factores globales que evolucionan conjuntamente según un VAR(4) y un factor específico del grupo para cada uno de los 8 grupos que se describieron

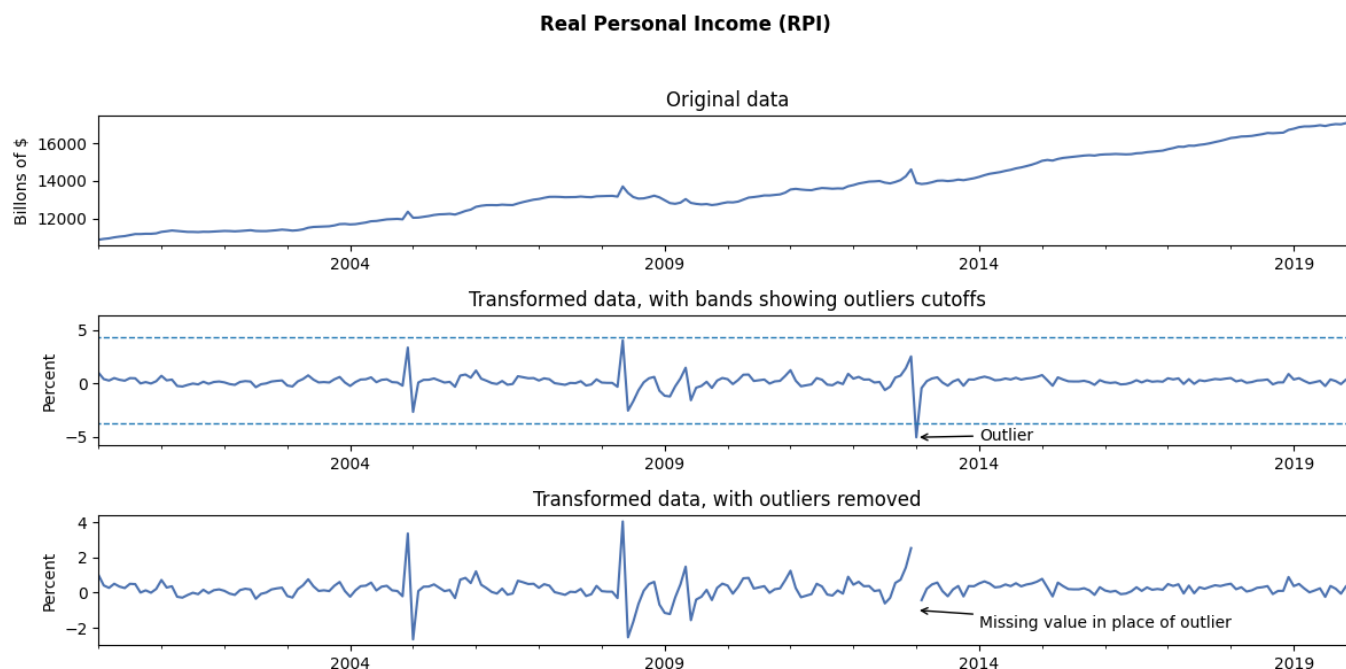


Figura 7. Ejemplo de transformación y la eliminación de atípicos de la variable RPI (‘Ingreso Personal Real’) durante el período de enero de 2000 a enero de 2020

anteriormente, cada uno evolucionando según un AR(1).

Con el modelo construido, el modelo puede ajustarse / los parámetros pueden estimarse mediante máxima verosimilitud a través del algoritmo EM.

Una vez obtenidos las cargas del modelo, por ejemplo, en la gráfica a continuación, el factor “Global.1” aumenta notablemente en 2009, después de la crisis financiera global. Sin embargo, muchas de las cargas de factores generados son negativas, por ejemplo, esto es cierto para las series de producción, consumo e ingresos. Por lo tanto, el aumento en el factor “Global.1” durante este período realmente implica una fuerte disminución en la producción, el consumo y los ingresos.

Una forma de examinar cómo se relacionan los factores con las variables observadas es calcular el poder explicativo que cada factor tiene para cada variable, mediante la regresión de cada variable en una constante más una o más de las estimaciones suavizadas de los factores y almacenando el valor resultante de R^2 , o coeficiente de determinación”. A continuación, se muestran las 10 principales variables observadas que son explicadas (individualmente) por cada uno de los dos factores globales.

Como se muestra en la 2 el primer factor global explica bien las series del mercado laboral, pero también incluye el PIB real y una medida de la volatilidad del mercado de valores (VXO). El segundo factor parece explicar en gran medida las variables relacionadas con la vivienda (de hecho, esto podría ser un argumento para eliminar el factor específico del grupo “Vivienda”).

Uno de los beneficios de estos modelos es que podemos utilizar la dinámica de los factores para producir pronósticos de cualquiera de las variables observadas. A continuación, se muestra un ejemplo del método de forecasting a dos cuatrimestres para las primeras cinco variables observadas.

A continuación en la figura 10, se muestran predicciones dentro de la muestra y pronósticos fuera de la muestra para algunos diferenciales entre los valores del Tesoro y la tasa de fondos federales.

Aunque los datos originales en los conjuntos de datos FRED-MD/QD para el PIB real están en “Billones de Dólares Encadenados de 2012”, esta variable se transformó a la tasa de crecimiento trimestral anualizada (cambio porcentual) para su inclusión en el modelo. De manera similar, la Tasa de Desempleo Civil originalmente estaba en “Porcentaje”, pero se transformó en el cambio de 1 mes (primera diferencia) para su inclusión en el modelo. Debido a que los datos transformados se proporcionaron al modelo, los métodos de predicción y pronóstico producirán predicciones y pronósticos en el espacio transformado. A continuación, en las figuras 11 y 12 se representan de manera gráfica los pronósticos directamente del modelo asociado con el PIB real y la tasa de desempleo.

REFERENCIAS

- [1] Spearman, C. (1904). “General Intelligence,” Objectively Determined and Measured. *The American Journal of Psychology*, 15(2), 201–292. <https://doi.org/10.2307/1412107>
- [2] Geweke, J. (1977). The dynamic factor analysis of economic time series. *Latent variables in socio-economic models*.
- [3] Sargent, T. J., & 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.

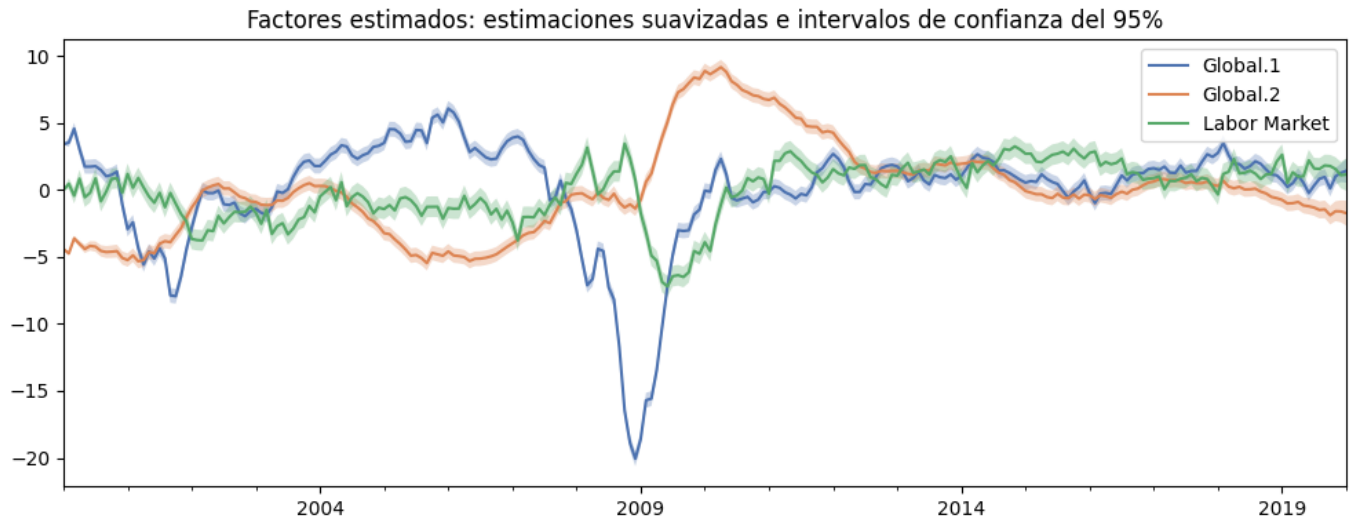


Figura 8. Factores estimados y su comparación con el mercado laboral

Tabla 2. Top diez variables explicadas por Global.1 (izquierda); Top diez variables explicadas por Global.2 (derecha)

Variable	R^2	Variable	R^2
All Employees: Total nonfarm	0.74	Housing Starts, South	0.66
All Employees: Goods-Producing Industries	0.72	New Private Housing Permits, South (SAAR)	0.66
All Employees: Durable goods	0.64	Housing Starts: Total New Privately Owned	0.65
All Employees: Wholesale Trade	0.64	New Private Housing Permits (SAAR)	0.65
All Employees: Manufacturing	0.63	New Private Housing Permits, West (SAAR)	0.64
All Employees: Service-Providing Industries	0.63	Housing Starts, West	0.64
All Employees: Trade, Transportation & Utilities	0.60	New Private Housing Permits, Midwest (SAAR)	0.55
Real Gross Domestic Product, 3 Decimal (Billio...	0.53	New Private Housing Permits, Northeast (SAAR)	0.55
All Employees: Construction	0.53	Housing Starts, Northeast	0.52
Total Business Inventories	0.44	Housing Starts, Midwest	0.51

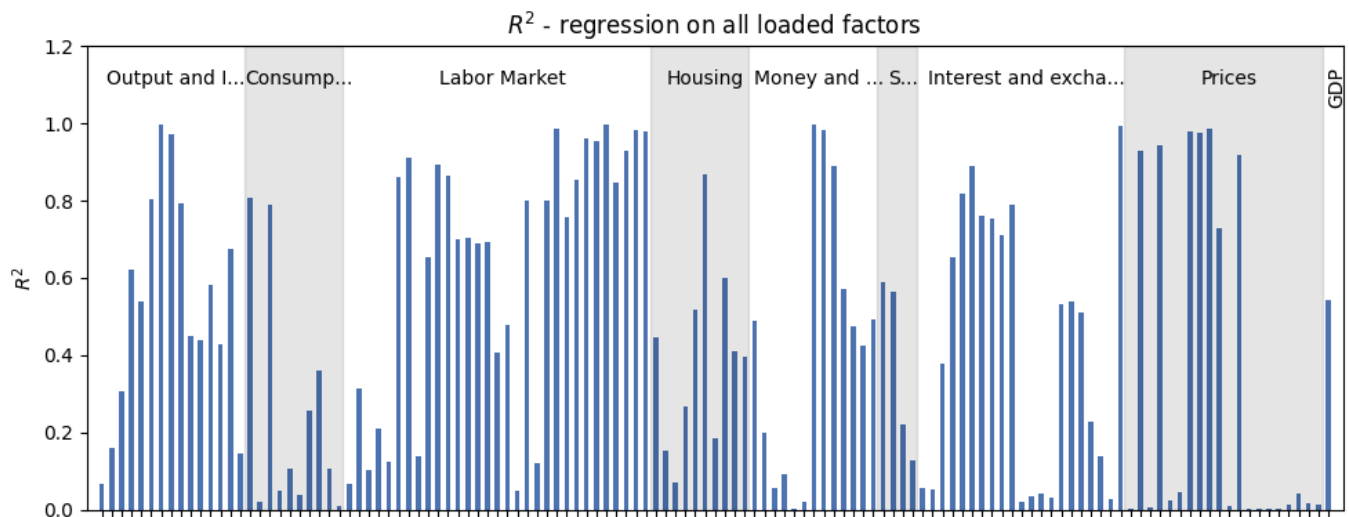


Figura 9. Valor explicativo general para una variable dada de todos los factores en los que carga la variable, para facilitar la identificación de patrones, se agregó sombreado y etiquetas para identificar los diferentes grupos de variables, así como nuestra única variable trimestral, el PIB.

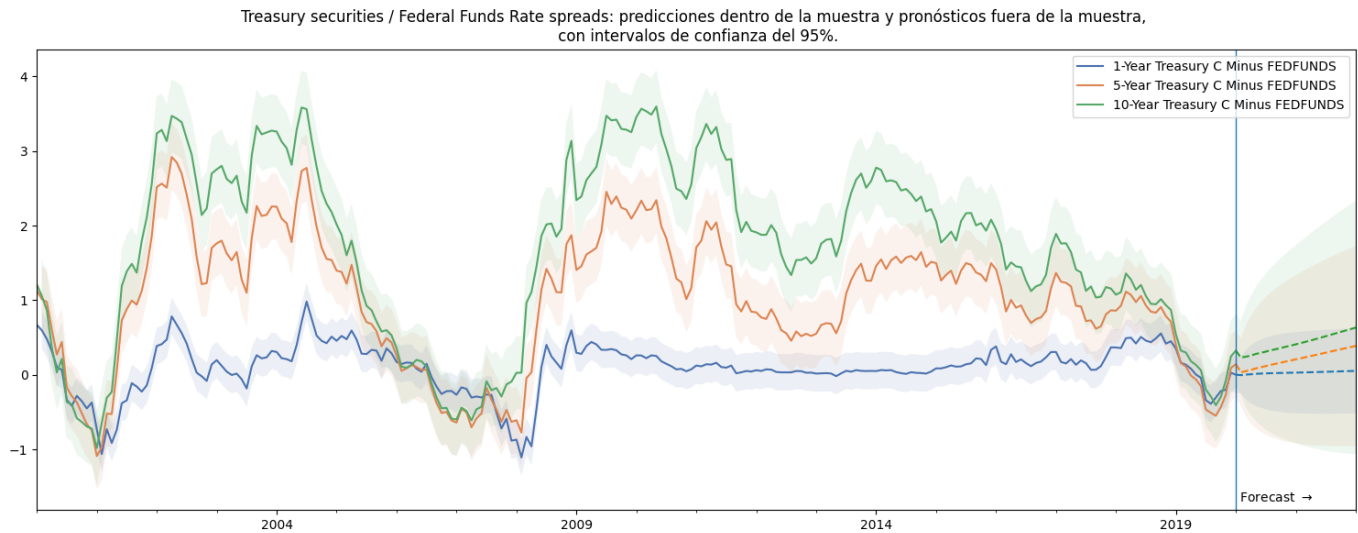


Figura 10. Pronósticos dentro y fuera de la muestra para algunos diferenciales entre los valores del Tesoro y la tasa de fondos federales

Datos y pronósticos (vintage de febrero de 2020), escala transformada.

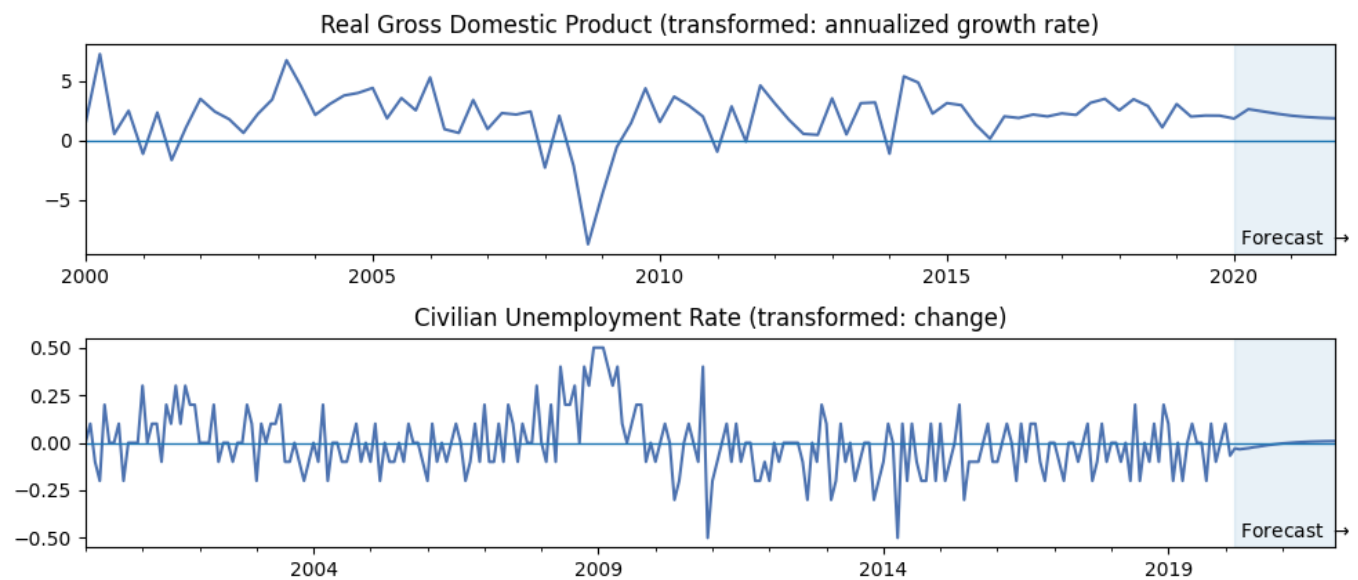


Figura 11. Pronósticos directamente del modelo asociado con el PIB real y la tasa de desempleo con datos transformados

Datos y pronósticos (vintage de febrero de 2020), escala original.

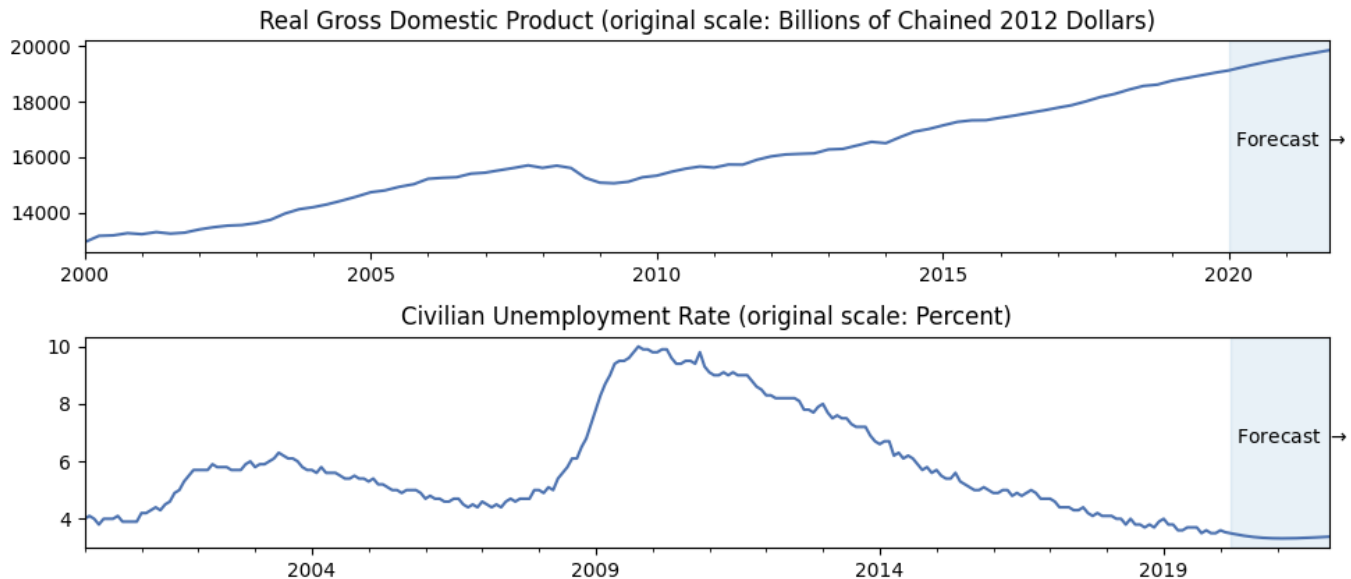


Figura 12. Pronósticos directamente del modelo asociado con el PIB real y la tasa de desempleo con datos originales

Variable	2020-02	2020-03
Real Personal Income	0.221031	0.268978
RPI ex transfer receipts	0.234809	0.266359
Real P. consumption expenditures	0.294328	0.230914
Real Manu. Trade Industries Sales	0.222638	0.230408
Retail Food Services Sales	0.473264	0.405351

Tabla 3. Pronósticos a dos cuatrimestres de las primeras cinco variables

- [4] Chamberlain, G. (1983). Funds, factors, and diversification in arbitrage pricing models. *Econometrica: Journal of the Econometric Society*, 1305-1323.
- [5] Chamberlain, G., & Rothschild, M. (1982). Arbitrage, factor structure, and mean-variance analysis on large asset markets.
- [6] Bai, J. (2003). Inferential theory for factor models of large dimensions. *Econometrica*, 71(1), 135-171.
- [7] Stock, J. H., & Watson, M. W. (2002). Forecasting using principal components from a large number of predictors. *Journal of the American statistical association*, 97(460), 1167-1179.
- [8] Stock, J. H., & Watson, M. W. (2002). Macroeconomic forecasting using diffusion indexes. *Journal of Business & Economic Statistics*, 20(2), 147-162.
- [9] Bai, J., & Ng, S. (2002). Determining the number of factors in approximate factor models. *Econometrica*, 70(1), 191-221.
- [10] Forni, M., Hallin, M., Lippi, M., & Reichlin, L. (2000). The generalized dynamic-factor model: Identification and estimation. *Review of Economics and statistics*, 82(4), 540-554.
- [11] Barigozzi, M., & Hallin, M. (2023). Dynamic Factor Models: a Genealogy. *arXiv preprint arXiv:2310.17278*.
- [12] Marotta, F., & Mumtaz, H. (2023). Vulnerability to climate change: evidence from a dynamic factor model. *Smith School Working Paper 23-06*, University of Oxford.
- [13] Peracchi, F., & Rossetti, C. (2022). A nonlinear dynamic factor model of health and medical treatment. *Health Economics*, 31(6), 1046-1066.
- [14] Molenaar, P. C., & Ram, N. (2009). Advances in dynamic factor analysis of psychological processes. In *Dynamic process methodology in the social and developmental sciences* (pp. 255-268). New York, NY: Springer US.
- [15] Choi, I. (2012). Efficient estimation of factor models. *Econometric Theory*, 28(2):274-308.
- [16] Stock, J. H. and Watson, M. W. (2005). Implications of dynamic factor models for var analysis. Technical report, National Bureau of Economic Research.
- [17] Bai, J. and Liao, Y. (2013). Statistical inferences using large estimated covariances for panel data and factor models. *arXiv preprint arXiv:1307.2662*.
- [18] Ahn, S. C. and Horenstein, A. R. (2013). Eigenvalue ratio test for the number of factors. *Econometrica*, 81(3):1203-1227.



- [19] Onatski, A. (2009). Testing hypotheses about the number of factors in large factor models. *Econometrica*, 77(5):1447–1479.
- [20] Onatski, A. (2010). Determining the number of factors from empirical distribution of eigenvalues. *The Review of Economics and Statistics*, 92(4):1004–1016.
- [21] Kapetanios, G., and M. Marcellino (2008), “Factor-GMM Estimation with Large Sets of Possibly Weak Instruments,” manuscript, EUI
- [22] Bai, J. and S. Ng (2010), “Instrumental Variable Estimation in a Data-Rich Environment,” Forthcoming, *Econometric Theory*
- [23] Kloeck, T., and L.B.M. Mennes (1960), “Simultaneous Equations Estimation Based on Principal Components of Predetermined Variables,” *Econometrica*, 28, 45-61.
- [24] Amemiya, T. (1966), “On the Use of Principal Components of Independent Variables in Two-Stage Least-Squares Estimation,” *International Economic Review* 7, 283-303.
- [25] Doz, C., Giannone, D., & Reichlin, L. (2011). A two-step estimator for large approximate dynamic factor models based on Kalman filtering. *Journal of Econometrics*, 164(1), 188-205.
- [26] Bai, J. and Ng, S. (2006). Confidence intervals for diffusion index forecasts and inference for factor-augmented regressions. *Econometrica*, 74(4):1133–1150.
- [27] Fulton, C. (8 diciembre 2023). Large dynamic factor models, forecasting, and nowcasting. Recuperado de https://www.chadfulton.com/topics/statespace_large_dynamic_factor_models.html