Centro de Investigación en Matemáticas

Unidad Monterrey

Pronósticos vía VAR-PLS

T E S I N A

QUE PARA OPTAR POR EL GRADO DE:

Maestro en Cómputo Estadístico

PRESENTA:

José Antonio García Ramírez

ASESOR DE TESIS:
Dra. Graciela Ma. de los Dolores González Farías
CO-ASESOR DE TESIS:
Dr. Francisco de Jesús Corona

Monterrey, Nuevo León, 2019

RESUMEN

En la actualidad los pronósticos de series de tiempo proporcionan información que son de ayuda en la toma de decisiones. Este tipo de dato posee peculiaridades pues considera la dimensión temporal, que aunado a la interacción con otras variables enriquecen el modelado estadístico.

Algunas de las técnicas para realizar pronósticos precisos requieren de técnicas computacionales desarrolladas recientemente, entre las cuales se encuentran: técnicas de reducción de dimensionalidad, los métodos Monte Carlo y los métodos de remuestreo; que permiten considerar al mismo tiempo más variables y hace factible, en términos de complejidad computacional, la estimación del pronóstico.

Con los resultados y la metodología llevada a cabo se generó en México un antecedente computacional y estadístico que permite efectuar pronósticos altamente precisos haciendo uso de diversas variables econométricas.

A mis padres Claudia y Nicolás, por su apoyo incondicional y su enorme esfuerzo al educarme, y a mis hermanos quienes alegran mi vida con sus cariños. A Dios por mi familia y esta vida.

Agradecimientos

Agradezco a la vida por permitirme conocer a la Mtra. Alma Luz quien me indujo en la educación básica y preparatoria hacia las matemáticas. Luego en la facultad de ciencias al Mtro. Lara Aparicio por mostrarme de qué se trata la matemática aplicada. Y más recientemente, desde Acatlán y durante la maestría, doy gracias al Dr. Arturo Erdely por mostrarme que la estadística y el cómputo convergen aunado a su apoyo y amistad.

En la bella ciudad de Monterrey quiero reconocer a mi asesora y a mi tutor, la Dra. Graciela González Farías y el Dr. Rodrigo Macías, cuya afectuosa y rigurosa guía me permitieron concretar mis estudios de posgrado incluyendo la participación en varios eventos dentro y fuera del país.

Gracias a la unidad Monterrey de CIMAT y quienes trabajan en ella, por su esfuerzo que me permitió seguir con alegría y entusiasmo este desafío.

Quiero reconocer el apoyo de mi país y en particular a CONACYT por la beca que me otorgó para la realización de este programa de maestría, además al programa **XXX** cuya ayuda me permitió asistir a la universidad de Washington en el verano del 2018.

Finalmente, pero no menos importante, quiero agradecer a mis amigos y compañeros de licenciatura y maestría cuya alegría me permitió pasar los días difíciles: Cesar, Eduardo, Edison, Erick, Hairo, ...

Contenido

CAPÍTULO 1

Introducción

Durante el siglo pasado y el presente muchas organizaciones tendieron a desarrollar e incorporar un gran número de sistemas automatizados para registrar y almacenar con precisión grandes cantidades de datos. Por ejemplo, el comercio electrónico permite registrar transacciones diversas en tiempo real; los supermercados llevan inventarios que se actualizan con cada compra o recepción de productos, González V. (2011).

Como consecuencia las organizaciones tienen una nueva meta: utilizar estos conjuntos de datos para apoyar y mejorar tanto la investigación como la toma de decisiones. Para ello una vía es utilizar modelos matemáticos que reflejen de manera aproximada el fenómeno que se analiza.

Una de las actividades más importantes que hace uso de un conjunto de datos es el de pronosticar, tomando en cuenta aspectos importantes como: los cambios en tendencias o variaciones debidas a fenómenos repetitivos como las estaciones del año o las quincenas, de tal modo que los pronósticos sean más precisos y por ende más útiles. Por lo general, al tomar decisiones es necesario utilizar modelos estocásticos, ya que se requiere de predicciones sobre un futuro más o menos incierto. Estos métodos cuantitativos permiten medir, inclusive, la confianza que puede tenerse en el pronóstico.

En México, como todos los demas países, donde toda suerte de decisiones políticas afecta a los procesos económicos y sociales, se hace más patente la necesidad de apoyarse en metodologías de pronósticos confiables, precisos y eficientes.

introducción econométrica de la tesis de Banxico

Objetivo del trabajo

Incluir objetivo principal y objetivos secundarios.

Motivación

A continuación, se reseñan los trabajos que motivaron la metodología VAR-PLS, la cual será detallada en el tercer capítulo.

Phillip Hans Franses en el 2006, Frances (2006), propuso una metodología para realizar pronósticos conjuntos de manera óptima a través de una representación autorregresiva de orden p, esto para h pasos adelante. Lo relevante de la metodología fue plantear un modelo denominado *Mínimos Cuadrados Parciales (PLS) Autorregresivo* ($PLSAR(h,p)$).

Este modelo se encuentra situado entre un $AR(p)$ que pronosticó todos los h pasos adelante y también para diferentes modelos AR en diferentes horizontes, de tal forma que formuló esos tres modelos para realizar h pronósticos.

La problemática computacional implicada en estos tipos de estudios está relacionada con la complejidad (volumen de datos y alta dimensión), la cual dificulta su captura, gestión, procesamiento, análisis e interpretación de los resultados. Por lo tanto, en forma específica, esta tesina tiene como **objetivo final efectuar pronósticos multivariados precisos**.

Para conseguir estos objetivos se presenta una descripción breve de lo que es el pronóstico de series de tiempo y los conceptos que esta metodología (VAR-PLS) requiere.

Hablar sobre el aspecto econométrico del INPC y el tipo de cambio, citar a INEGI y a Banxico y los problemas entre las variables 'verdaderas' -desde el punto de vista económico- y las variables observables[1].

En la actualidad, la dinámica económica en el país ... **impacto económico y geográfico**
Por tanto, en la conceptualización del problema, existen tres puntos a considerar:

- **El tiempo de procesamiento**, el cual suele ser elevado al tener que analizar grandes cantidades de datos, resultado de utilizar varias series de tiempo.
- **Las variables utilizadas**, son de difícil determinación **econometrica** e influyen considerablemente en el resultado final.
- **La evaluación** final de los datos y su interpretación.

El cómputo estadístico es un campo que permitirá, aunado con especialistas en el área a donde se enfoque, a la construcción de nuevos procesos de analisís de datos de

[1] Tema que se revisa en el tercer capítulo.

alta complejidad en función a los objetivos propuestos.

Así pues, este trabajo de investigación tiene como objetivo principal dar pronósticos contemplando variables **macroeconómicas** por medio de herramientas estadísticas computacionales.

El contenido de esta investigación se encuentra dividido en los siguientes capítulos:

Fundamentos teóricos. Se introducen los conceptos técnicos necesarios para el planteamiento y desarrollo de la metodología VAR-PLS.

Metodología. Se describen los métodos estadísticos y computacionales del VAR-PLS. A la par se incluye información referente a los datos, su obtención y una sutil justificación de la selección de variables a utilizar.[1]

Resultados. Se presenta los resultados del análisis, así como su implementación y se contrasta con **los resultados de otra metodología**[2].

Conclusiones y futuros trabajos. Se discuten los resultados y se realizan las conclusiones y recomendaciones del proyecto e investigación futura. [3]

[1] Elementos económicos y econométricos.

[2] Se liga con los resultados obtenidos en la estancia

[3] Series de alta latencia como las criptomonedas y acciones, imputación de series y matrices esparcidas.

CAPÍTULO 2

Fundamentos teóricos

En este capítulo se abordan las bases de los elementos estadísticos y computacionales utilizados para el desarrollo de la metodología VAR-PLS que se plantea en el tercer capítulo. A grandes rasgos el capítulo se compone de cuatro secciones:

En la sección 2.1 se revisan brevemente los conceptos fundamentales concernientes a las series de tiempo univariadas: definición probabilística, los supuestos que permiten su estudio estadístico, los modelos clásicos AR y su relación con las raíces unitarias.

Le sigue la sección 2.2; la cual inspecciona el modelo VAR partiendo de su reciente historia y uso dentro del modelado econométrico hasta su formulación estadística, continuando con el importante concepto de cointegración, las dinámicas a corto y largo plazo y su relación con el modelo de corrección de errores (MCE).

Posteriormente, en la sección 2.3 se describen algunos métodos de reducción de dimensionalidad y de regresión supervisada, que no deben confundirse con los métodos de selección de variables, que forman la parte perteneciente al pronóstico conjunto de la metodología VAR-PLS. Esta sección considera al análisis de componentes principales como introducción y a la vez como diferenciador del método de PLS.

Las tres secciones anteriores son importantes pues forman un breve resumen sobre estos tópicos, ya que el modelo VAR-PLS que se desarrollada más adelante se basa en estas ideas.

Finalmente, en la sección 2.4 se reseñan el método Bootstrap y algunas técnicas de simulación como lo son los métodos Monte Carlo; los cuales serán de utilidad, respectivamente, en la construcción de los intervalos de confianza y en la evaluación empírica de los pronósticos de la metodología propuesta.

2.1 Series de tiempo

Llamamos serie de tiempo a un conjunto de observaciones x_t, donde se observa y registra cada una de ellas en un instante de tiempo t y se asume que las observaciones son igualmente espaciadas.

Las series de tiempo pueden ser discretas en el tiempo, por ejemplo la cantidad de personas que usan diariamente el sistema de transporte colectivo; o bien continuas, como la temperatura que registra un sensor dentro de un horno a lo largo del día. En este trabajo se trabaja con series de tiempo discretas en su dominio.

Si bien los datos son el registro de lo que se puede observar de una serie de tiempo, se requiere de un modelo matemático para representar el proceso generador de los mismos, poder realizar inferencia al respecto y después de desarrollar un modelo satisfactorio podremos realizar predicciones, también llamados pronósticos, de los valores futuros[1] de la serie.

Formalmente cada observación x_t es la realización de una variable aleatoria X_t, por lo que el conjunto de datos que forman la serie de tiempo es una realización de la familia de variables aleatorias $\{X_t, t \in T_0\}$, a su vez esta familia de variables aleatorias forman parte de un proceso estocástico $\{X_t, t \in T\}$, donde $T_0 \subset T$, definido sobre un espacio de probabilidad $(\Omega, \mathcal{F}, P)$. En el contexto de series de tiempo llamamos índice a T y suele ser un subconjunto de $\Re$.

Recordando la definición de variable aleatoria se debe notar que para un instante de tiempo fijo $t \in T$, X_t es una función $X_t(.)$ en el conjunto de eventos Ω y por otro lado al fijar un evento $w \in \Omega$, el caso de interés en este trabajo, $X_.(w)$ es una función de T.

La idea a destacar de los conceptos anteriores es que una serie de tiempo se refiere tanto a los datos observados como a la realización de un proceso estocástico, de manera general el teorema de Kolmogorov asegura que bajo ciertas condiciones todo proceso estocástico posee una función de distribución, Brockwell and Davis (1986).

2.1.1 Estacionariedad y estacionariedad estricta

Si una serie de tiempo $\{X_t, t \in T\}$ tiene varianza finita para todo $t \in T$, entonces su función de autocovarianza $\gamma_x(.,.)$ se define como:

$$\gamma_x(r,s) = Cov(X_r, X_s) = E\left[(X_r - E(X_r))(X_s - E(X_s))\right], \quad \text{con} \quad r,s \in T \tag{2.1}$$

[1] También es posible realizar predicciones en tiempo pasado, pero en la práctica el conocimiento a futuro presenta mayor aplicación.

Se dice que una serie de tiempo es estacionaria si cumple las siguientes tres propiedades:

1. $E(|X_t|^2) < \infty$, para todo $t \in \mathbb{Z}$
2. $E(X_t) = m$, para todo $t \in \mathbb{Z}$
3. $\gamma_x(r, s) = \gamma_x(r + t, s + t)$, para todo $r, s, t \in \mathbb{Z}$

Las primeras dos condiciones anteriores establecen que la varianza y el valor esperado de la serie de tiempo son iguales en cualquier instante de tiempo t y la tercera condición define la estacionariedad débil, como se conoce en la literatura, que establece que la autocovarianza de una serie de tiempo en dos momentos diferentes es función solo de la longitud del intervalo entre ellas y no depende de los instantes en que se observa.

Como consecuencia se tiene que si una serie de tiempo es estacionaria entonces $\gamma_x(r, s) = \gamma_x(r - s, 0)$, $\forall t, s \in \mathbb{Z}$; por lo cual su función de autocovarianza se simplifica a la forma:

$$\gamma_x(h) = \gamma_x(h, 0) = Cov(X_{t+h}, X_t), \quad \forall t, h \in \mathbb{Z}$$

Y su función de autocorrelación con rezago h se define como:

$$\rho_x(h) = \gamma_x(h)/\gamma_x(0) = Corr(X_{t+h}, X_t), \quad \forall t, h \in \mathbb{Z}$$

Por otro lado la estacionariedad estricta se tiene cuando las distribuciones conjuntas de $(X_{t_1}, \ldots, X_{t_k})'$ y $(X_{t_1+h}, \ldots, X_{t_k+h})'$ son iguales, esto puede interpretarse como que las realizaciones de una serie de tiempo en dos intervalos de tiempo de la misma longitud poseen características estadísticas similares. Como puede consultarse en Brockwell and Davis (1986) la estacionariedad estricta implica a la débil y el inverso solo es válido en el caso de la distribución normal multivariada. Para fines prácticos la estacionariedad débil es más fácil de establecer que su versión estricta, por lo que se trabaja con ella. En lo siguiente estacionariedad se refiere a estacionariedad débil.

Es común descomponer a una serie de tiempo X_t de la siguiente manera:

$$X_t = m_t + s_t + \epsilon_t$$

Donde m_t es la componente de tendencia; s_t es la componente estacional, que puede contener ciclos de diversas longitudes, y ϵ_t es una componente aleatoria y estacionaria.

Para lograr la descomposición anterior es usual recurrir a transformaciones de la serie, aplicando por ejemplo la función ln(.), o bien diferenciando la serie[1] para aislar la componente ϵ_t y modelarla como un proceso estacionario. Existen diferentes estrategias, en las cuales el presente trabajo no profundiza, para extraer la componente de

[1] Sobrediferenciar una serie de tiempo puede incrementar la varianza de la componente aleatoria de manera que esta deje de ser estacionaria, Chan (2010).

tendencia, por estimación de mínimos cuadrados o bien la implementación del algoritmo X-13ARIMA-SEATS, Sax and Eddelbuettel (2018), para remover la estacionalidad.

Las estimaciones muestrales de la función de autocovarianza y de autocorrelación se definen, respectivamente, de la siguiente forma:

$$\hat{\gamma}(h) = n^{-1} \sum_{j=1}^{n-h} (x_{j+h} - \bar{x})(x_j - \bar{x}), \quad 0 \leq h \leq n$$

$$\hat{\rho}(h) = \hat{\gamma}(h)/\hat{\gamma}(0), \quad |h| < n$$

Donde $\bar{x}$ es la media muestral, $\bar{x} = n^{-1} \sum_{j=1}^{n} x_j$.

2.1.1.1 Ergodicidad

Un concepto de especial interés en el análisis de series de tiempo es el de ergodicidad. Se dice que un proceso estocástico es ergódico si la media muestral de una realización del proceso converge al subyacente parámetro del proceso.

Para procesos ergódicos no se requiere observar varias realizaciones independientes del proceso para determinar su media y momentos de orden superior, pues la convergencia de la media muestral garantiza la existencia del primer y segundo momento que es compartido por todas las realizaciones del proceso estocástico, Chan (2010), del que además se puede estimar la autocovarianza.

2.1.2 Modelos AR

Una categoría de modelos, ampliamente usados en el análisis de series de tiempo, son los modelos autorregresivos AR[1]. Estos modelos tienen una interpretación sencilla y se parecen a los modelos de regresión lineal tradicionales cuando se reemplaza al predictor por su valor anterior rezagado.

Sea B al operador de rezago de tal forma que si $\{x_1, \ldots, x_t\}$ es una serie de tiempo, se tiene que $BX_t = X_{t-1}$ y de manera análoga al iterar n veces el operador B obtenemos que $B^n X_t = X_{t-n}$.

Formalmente un modelo $AR(p)$, autorregresivo de orden p, se puede escribir como $\phi(B)X_t = \epsilon_t$, donde $\phi(B) = (1 - \phi_1 B - \cdots - \phi_p B^p)$, a fin de que:

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

Donde $\{X_t\}$ es estacionario. Análogamente se dice que una serie de tiempo $\{X_t\}$ es un proceso $AR(p)$ con media μ si $\{X_t - \mu\}$ es un proceso $AR(p)$.

[1]Del inglés Autoregressive Model

Una cualidad importante que algunos modelos AR poseen es la causalidad. Se dice que un proceso AR es causal si su valor en el tiempo t depende sólo de los valores observados anteriores, $t^* < t$, y no de valores futuros, es decir que existen constantes $\{\psi_i\}$ con $\sum_{i=0}^{\infty} |\psi_i| < \infty$ tales que $X_t = \sum_{i=0}^{\infty} \psi_i \epsilon_{t-i}$. De manera equivalente un proceso $AR(p)$ es causal si las raíces del polinomio característico $\phi(z) = 1 - \phi_1 z - \cdots - \phi_p z^p$ se encuentran fuera del círculo unitario $\{z : |z| > 1\}$[1].

Los parámetros $\{\phi_i\}$ de un modelo $AR(p)$ pueden estimarse por mínimos cuadrados o bien por máxima verosimilitud, lo cual requiere de suponer una distribución normal a la componente aleatoria ϵ_t. Un aspecto fundamental en esta estimación es la determinación del parámetro p que indica el orden del rezago del modelo AR. Para la determinación de este parámetro se pueden seguir dos métodos[2]: el error de predicción final (FPE) y el criterio de información de Akaike (AIC).

Para ilustrar el FPE, considérese una realización $X = (X_1, \ldots, X_n)$ de un modelo $AR(p)$ con $p < n$. Si $\hat{\phi}_1, \ldots, \hat{\phi}_p$ son las estimaciones de máxima verosimilitud a partir de X entonces se puede estimar la varianza $\hat{\sigma}^2$, de los errores $\hat{\epsilon}_t$, lo cual permite definir:

$$FPE = \hat{\sigma}^2 \left(\frac{n+p}{n-p} \right)$$

De donde se busca el valor de p que minimice el FPE.

2.1.3 Raíces unitarias

Hasta este punto se ha hablado de series de tiempo estacionarias, pero en la práctica se encuentran diferentes tipos de no-estacionariedad. La no-estacionariedad puede deberse a cambios sistemáticos en la media o en la varianza. Cuando la no-estacionariedad se debe a cambios en la varianza esta puede controlarse por medio de aplicar una transformación de los datos como la función ln(.) o bien cuando se observa heteroscedasticidad, bajo ciertos supuestos, el proceso puede modelarse como un ARCH[3], Hamilton (1994).

Cuando la no-estacionariedad se debe a cambios en la media se presentan las raíces unitarias. Consideremos la prueba para el coeficiente del $AR(1)$, $H : \alpha = 1$ para el modelo:

$$X_t = \beta_1 + X_{t-1} + \epsilon_t$$

Para ilustrar la idea consideremos $\beta_1 = 0$ en la ecuación anterior, entonces bajo H,$\{X_t\}$ es una caminata aleatoria y las pruebas estadísticas para este tipo de modelo

[1] En esta notación z es un número complejo.

[2] Una revisión a detalle de la selección del orden p puede consultarse en Chan (2010)

[3] Del inglés Autoregressive Conditional Heteroskedasticity

son conocidas en la literatura de series de tiempo y econométrica como pruebas de raíz unitaria. Si se supone una distribución normal e independencia en la componente estocástica ϵ_t la estimación por mínimos cuadrados $\hat{\alpha}$ para α es:

$$\hat{\alpha} = \frac{\sum_{t=1}^{n} Y_t Y_{t-1}}{\sum_{t=1}^{n} Y_{t-1}^2}$$

En particular

$$n(\hat{\alpha} - 1) = \frac{(1/n)\sum_{t=1}^{n} Y_{t-1}\epsilon_t}{(1/n^2)\sum_{t=1}^{n} Y_{t-1}^2} \tag{2.2}$$

Puede demostrarse, Chan (2010), que el lado izquierdo de la ecuación anterior tiene una convergencia casi segura a una expresión en términos del movimiento Browniano. El resultado anterior provee la prueba de Dickey-Fuller para raíces unitarias donde la restricción de $\beta = 0$ puede eliminarse y se conoce como la prueba de Dickey-Fuller aumentada.

La presencia y determinación de raíces unitarias en series de tiempo de variables macroeconómicas es un tema ampliamente estudiado; en el segundo capítulo de Juselius (2007) se puede encontrar un ejemplo que involucra a la inflación anual de Dinamarca observada en los diferentes intervalos: 1901-1992, 1945-1992 y 1975-1992; donde para los dos primeros se tiene una serie estacionaria y en el tercero -subconjunto del segundo- la estacionariedad se pierde. El hecho de que la inflación sea estacionaria o no es un tema que ha tenido mucho debate, algunos basados en la interpretación de cambios estructurales (económicos) que exhibe la raíz unitaria. En este contexto el hecho de que el coeficiente α se acerque a la unidad puede no ser significativo debido a un reducido tamaño de muestra. En los casos en que el tamaño de muestra es grande y la serie presenta una raíz unitaria con orden de integración igual a la unidad, desde un punto de vista empírico se dice que sus propiedades se mantienen en el largo plazo.

Lo anterior permite concluir con la reseña de las series de tiempo univariadas al definir el orden de integración. Consideremos el proceso $X_t = \sum_{i=1}^{n} \epsilon_i$, donde ϵ_t es una secuencia de variables aleatorias con distribución normal descorrelacionadas con media cero y varianza σ^2, usualmente llamado ruido blanco. Entonces X_t tiene varianza $t\sigma^2$ así que X_t es no-estacionaria y su orden de integración es 1.

Decimos que el proceso $X_t = \sum_{i=1}^{\infty} \phi_i \epsilon_i$ es integrado de orden cero, denotado como $I(0)$, si ϵ_t es ruido blanco y $\sum_{i=1}^{\infty} \phi_i \neq 0$. De manera análoga el proceso $\{X_t\}$ es integrado de orden 1, si $\Delta X_t = (1 - B)X_t = X_t - X_{t-1}$ es $I(0)$. De manera general un proceso X_t con d raíces unitarias se dice que tiene orden de integración d, $I(d)$.

2.2 Modelos VAR

Se dice que una serie de tiempo k-variada es un proceso estocástico que contiene vectores de dimensión k, $(X_{t1}, X_{t2}, \ldots, X_{tk})'$ observados en los tiempos t (usualmente $t = 0, 1, 2, \ldots$). Las componentes de la serie $\{X_{ti}\}$ pueden ser estudiadas independientemente como una serie de tiempo univariada, cada una caracterizada por su media y función de autocovarianza. Tal aproximación falla al considerar posible dependencia entre componentes, y tal dependencia cruzada puede ser de gran importancia para predecir valores futuros de cada componente.

Considérese la serie de vectores aleatorios $\boldsymbol{X_t} = (X_{t1}, \ldots, X_{tk})'$, se define su vector de medias como:

$$\mu_t = E(X_t) = (E(X_{t1}), \ldots, E(X_{tk}))'$$

Y matriz de covarianza:

$$Cov(X_{t+h,i}, X_{t,i}) = \begin{pmatrix} \gamma_{11}(t+h,t) & \ldots & \gamma_{1k}(t+h,t) \\ \vdots & & \vdots \\ \gamma_{k1}(t+h,t) & \ldots & \gamma_{kk}(t+h,t) \end{pmatrix}$$

Donde

$$\gamma_{ij}(t+h,t) = Cov(X_{t+h,i}, X_{t,j})$$

En notación matricial:

$$\Gamma(t+h,t) = E((X_{t+h} - \mu_{t+h})(X_t - \mu_t))$$

Se dice que la serie $\boldsymbol{X_t}$ es estacionaria si los momentos $\boldsymbol{\mu_t}$ y $\boldsymbol{\Gamma}(t+h,t)$ son ambos independientes de t, en ese caso se usa la notación

$$\mu = E(X_t)$$

Y

$$\Gamma(h) = Cov(X_{t+h}, X_t)$$

Los elementos en la diagonal de la matriz anterior son las autocovarianzas de las series univariadas $\{X_{ti}\}$, mientras que los elementos fuera de la diagonal son las covarianzas entre $X_{t+h,i}$ y $X_{t,j}$, con $i \neq j$. Nótese que $\gamma_{ij}(h) = \gamma_{ji}(-h)$. De manera correspondiente la matriz de correlación se define como:

$$R(h) = \begin{pmatrix} \rho_{11}(h) & \ldots & \rho_{1k}(h) \\ \vdots & & \vdots \\ \rho_{k1}(h) & \ldots & \rho_{kk}(h) \end{pmatrix}$$

Donde

$$\rho_{ij} = \gamma_{ij}(h)(\gamma_{ii}(0)\gamma_{jj}(0))^{-1/2}$$

Denotamos como $\{\boldsymbol{\epsilon_t}\} \sim i.i.d.\ (0,\ \boldsymbol{\Sigma})$ si $\{\boldsymbol{\epsilon_t}\}$ son independientes e idénticamente distribuidas con media $\boldsymbol{\mu}$=**0** y matriz de covarianza $\boldsymbol{\Sigma}$.

$\boldsymbol{X_t}$ es un proceso lineal si puede ser expresado como:

$$X_t = \sum_{j=-\infty}^{\infty} C_j \epsilon_{t-j}$$

Donde $\boldsymbol{\epsilon_t}$ es ruido blanco y $\{\boldsymbol{C_j}\}$ es una secuencia de matrices de tamaño $k \times k$ cuyas entradas son absolutamente sumables, es decir:

$$\sum_{j=-\infty}^{\infty} |C_j(i,l)| < \infty, \text{ con } i,l = 1,2,\ldots,k$$

Con lo anterior puede definirse el modelo de series de tiempo, vector autorregresivo de orden p, $VAR(p)$, como:

$$X_t = \nu + \Phi_1 X_{t-1} + \cdots + \Phi_p X_{t-p} + \epsilon_t \tag{2.3}$$

Con $\boldsymbol{\epsilon_t}$ ruido blanco, nótese que $\boldsymbol{\nu} = (\nu_1, \ldots, \nu_k)'$ es un vector fijo de términos de intercepto, permitiendo la posibilidad de media diferente de cero, $E(X_t)$.

La mayor parte de la teoría de series de tiempo univariadas puede extenderse al caso multivariado de manera natural, sin embargo surgen problemas. Por ejemplo la maldición de la dimensionalidad. Cuando el número de componentes en $\boldsymbol{X_t}$ se incrementa, el número de parámetros también crece. Por ejemplo para una serie de tiempo conteniendo 10 componentes, incluso en el sencillo caso de un modelo $AR(1)$ se tienen 100 parámetros libres.

Considérese un modelo $VAR(1)$:

$$X_t = \nu + \Phi_1 X_{t-1} + \epsilon_t$$

Si este modelo generador comienza en el tiempo $t = 1$ se tiene:

$$\begin{aligned}
X_1 &= \nu + \Phi_1 X_0 + \epsilon_1 \\
X_2 &= \nu + \Phi_1 X_1 + \epsilon_2 = \nu + \Phi_1(\nu + \Phi_1 X_0 + \epsilon_1) + \epsilon_2 \\
&= (I_K + \Phi_1)\nu + \Phi_1^2 X_0 + \Phi_1 \epsilon_1 + \epsilon_2 \\
&\vdots \\
X_t &= (I_K + \Phi_1 + \cdots + \Phi_1^{t-1})\nu + \Phi_1^t X_0 + \sum_{i=0}^{t-1} \Phi^i \epsilon_{t-i}
\end{aligned}$$

Continuando de esta manera podemos escribir el proceso $VAR(1)$ como:

$$X_t = \nu + \sum_{i=0}^{\infty} \Phi_1^i \epsilon_{t-i}$$

Donde el vector de medias de $\boldsymbol{X_t}$, $E(X_t) = \nu + \Phi_1\nu + \Phi_1^2\nu + \dots$ si Φ_1 tiene todos sus valores propios menor a 1 en módulo.

De esta manera se dice que un proceso $VAR(1)$ es estable si:

$$det(I_k - z\Phi_1) \neq 0, \text{ para } |z| \leq 1$$

Para un proceso general $VAR(p)$ la discusión anterior se puede extender reescribiendo cualquier $VAR(p)$ en forma de $VAR(1)$, específicamente podemos escribir:

$$\tilde{X}_t = \begin{pmatrix} X_t \\ X_{t-1} \\ \vdots \\ X_{t-p+1} \end{pmatrix}, \tilde{\nu} = \begin{pmatrix} \nu \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \tilde{\Phi} = \begin{pmatrix} \Phi_1 & \Phi_2 & \dots & \Phi_{p-1} & \Phi_p \\ I_k & 0 & \dots & 0 & 0 \\ 0 & I_k & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & I_k & 0 \end{pmatrix}, \tilde{\epsilon_t} = \begin{pmatrix} \epsilon_t \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

Donde $\boldsymbol{\tilde{X}_t}$, $\boldsymbol{\tilde{\nu}}$ y $\boldsymbol{\tilde{\epsilon}_t}$ son de dimensión $kp \times 1$ y $\boldsymbol{\tilde{\Phi}}$ es $kp \times kp$. De esta manera se tiene que $\boldsymbol{\tilde{X}_t}$ es estable si:

$$det(I_{kp} - z\tilde{\Phi}) \neq 0, \text{ para } |z| \leq 1$$

En la forma del modelo VAR (2.3) no existe solo un orden p correcto para el proceso. De hecho, si (2.3) es una forma correcta de escribir el proceso $\boldsymbol{X_t}$, lo mismo es cierto para:

$$X_t = \nu + \Phi_1 X_{t-1} + \dots + \Phi_p X_{t-p} + \Phi_{p+1} X_{t-p-1} + \epsilon_t$$

Con $\boldsymbol{\Phi_{p+1} = 0}$. En otras palabras si $\boldsymbol{X_t}$ es un proceso $VAR(p)$, en este sentido también es un $VAR(p+1)$. Es útil tener un único número que determine el orden del proceso $VAR(p)$. En lo siguiente se llamara a $\boldsymbol{X_t}$ un proceso $VAR(p)$ si $\boldsymbol{\Phi_p \neq 0}$ y $\boldsymbol{\Phi_i = 0}$ para $i > p$, de esta manera p es el orden posible más pequeño y define el orden del proceso $VAR(p)$. Esto es necesario, pues como se mencionó anteriormente en esta sección, el ajuste de modelos VAR con innecesarios órdenes de rezago incrementa el número de parámetros y puede disminuir la calidad de las estimaciones y por ende de los pronósticos, Lütkepohl (2006).

De manera análoga a los modelos AR, la determinación del orden de un modelo $VAR(p)$ es un aspecto muy importante a considerar. De manera general la determinación del orden p de un VAR se realiza por medio de una secuencia de pruebas donde las hipótesis son de la forma $H_0 : \Phi_m = 0$, si la hipótesis anterior se rechaza se prosigue

con la prueba $H_0 : \Phi_{m+1} = 0$ hasta que no sea rechazada la hipótesis nula.

Suponiendo normalidad para la componente estocástica de la serie multivariada $\boldsymbol{X_t}$, se puede recurrir al estadístico de Wald de la prueba de razón de verosimilitud para determinar el orden p, Lütkepohl (2006). Este enfoque se centra en determinar el proceso que genera los datos, sin embargo si nuestro objetivo es pronosticar existen otras maneras de determinar el orden de un modelo $VAR(p)$ como la elección de p por medio del valor que minimice el error cuadrático medio, el FPE o bien el criterio de Akaike (en sus versiones multivariadas), Lütkepohl (2006).

El enfoque anterior en la determinación del orden p requiere de dividir la muestra en dos partes: digamos que la primer parte consiste en las primeras l observaciones del vector $\boldsymbol{X_t}$ y el resto a las últimas $t - l$ observaciones, la primera será usada como conjunto de entrenamiento donde podemos utilizar la prueba de razón de verosimilitud para determinar un orden adecuado y con este orden predecir los siguientes $t-l$ valores del vector $\boldsymbol{X_t}$ de esta manera podemos utilizar los valores predichos y los reales para la elección del orden que mejor desempeño tenga al disminuir el criterio que hayamos elegido y fijado.

En la caracterización anterior de los modelos VAR, las matrices $\boldsymbol{\Phi_i}$ son constantes, aunque en la práctica es común que las series de tiempo univariadas presenten cambios estructurales que invaliden lo anterior. Por otro lado, después de que se ha determinado el orden p de un modelo VAR es importante validar los supuestos del modelo: la componente estocástica es ruido blanco y que los residuos están descorrelacionados (para diversos rezagos h). Para comprobar la normalidad de los residuos existen diversas pruebas como la prueba de Kolmogorov-Smirnov o la χ^2, Sheskin (2000), en cuanto a la descorrelación entre los residuos se puede emplear la prueba de Portmanteau o la prueba de multiplicadores de Lagrange, Lütkepohl (2006).

2.2.1 Contexto del uso de modelos VAR en econometría

Siguiendo a Juselius (2007) , los economistas frecuentemente formulan modelos bien especificados *económicamente* y aplican métodos estadísticos para estimar sus parámetros. En contraste, los estadísticos formulan modelos bien especificados *estadísticamente* para los datos y analizan el modelo estadístico para resolver las cuestiones económicas de interés. En el primer caso, la estadística es usada pasivamente como una herramienta para obtener algunas estimaciones, y en el segundo caso el modelo estadístico es usado activamente como medio de análisis del subyacente proceso generador del fenómeno en cuestión.

El principio general de analizar modelos estadísticos en macroeconomía fue introducido por Haavelmo en los 90's, Haavelmo (1944). El enfoque probabilístico de Haavelmo

a la econometría requiere de una formulación probabilística del proceso completo que genera los datos. Los aspectos computacionales involucrados en la especificación de los modelos estadísticos eran prohibitivos en los tiempos del trabajo de Haavelmo (1944), cuando incluso la estimación de una regresión múltiple era una tarea no trivial. El desarrollo del cómputo actual, hace factible las líneas adoptadas en Haavelmo (1944) en la econometría empírica. Es aquí donde los modelos VAR ofrecen ventajas como un marco de referencia general para dirigir las preguntas empíricas en (macro)-economía al mismo tiempo que se apegan al principio probabilístico general de Haavelmo.

Un aspecto importante mencionado en la pág. 5 en Haavelmo (1944) es el cómo las *verdaderas* variables -como funciones del tiempo- representan un ideal de mediciones precisas de la realidad, mientras que las variables definidas en teoría son las mediciones que se tienen si la realidad está de acuerdo con el modelo teórico. Es decir, que aún en el supuesto de que el modelo estadístico refleje adecuadamente la realidad, los datos pueden no medir directamente la variable en cuestión. Las mediciones disponibles de parte de las estadísticas oficiales, como es el caso de las variables que utilizaremos en los capítulos posteriores, pueden distar de las definiciones de las variables verdaderas. Cuestiones alrededor de las variables macroeconómicas que suelen reportarse como agregados[1] en unidades de tiempo como meses, trimestres, semestres, ... se relacionan con algunos principios generales para la modelación VAR con series no estacionarias.

2.2.2 Cointegración

Hasta este punto se ha hablado de modelos VAR estables, donde cada una de sus componentes es estacionaria. Pasemos ahora al caso general en donde el modelo VAR puede incluir series no-estacionarias.

De manera general un proceso $k-$dimensional $\boldsymbol{X_t}$ es llamado cointegrado de orden (d, b), escrito de manera breve, $\boldsymbol{X_t} \sim CI(d, b)$, si todas las componentes de $\boldsymbol{X_t}$ comparten el mismo orden de integración, $I(d)$, y existe una combinación lineal de las componentes de $\boldsymbol{X_t}$ tal que $\epsilon_t = \beta' X_t$ con $\beta = (\beta_1, \ldots, \beta_k)' \neq 0$ donde ϵ_t es $I(d-b)$.

Por ejemplo si todas las componente de $\boldsymbol{X_t}$ son $I(1)$ y $\beta' X_t$ es estacionario, $I(0)$, entonces $\boldsymbol{X_t} \sim CI(1, 1)$. El vector β es llamado vector de cointegración. Un proceso que contiene variables cointegradas es llamado cointegrado, así los procesos VAR con variables cointegradas con llamados modelos VAR cointegrados estudiados a profundidad en los últimos años, Juselius (2007). Los procesos cointegrados fueron introducidos por Granger en 1981, Granger (1981).

[1] Inclusive nuevas componentes pueden entrar al agregado o bien los cambios en tecnología y en la definición de los intervalos en que se realiza la medición pueden alterar el valor reportado.

El vector de cointegración anterior no es único, por ejemplo, pueden existir vectores de cointegración linealmente independientes que se pueden presentar en el caso de tener cuatro variables en un sistema donde las dos primeras están conectadas por una relación a largo plazo, al igual que las otras dos restantes. Entonces un vector de cointegración con ceros en sus dos últimas componentes refleja la primera relación, y un vector con ceros en sus dos primeras componentes refleja la segunda relación, además puede existir un vector de cointegración involucrando las cuatro variables.

La cointegración implica que ciertas combinaciones lineales de las componentes de la serie multivariada poseen un orden de integración menor al del proceso por sí mismo. Las variables cointegradas son accionadas por los mismos choques persistentes, o cambios en el tiempo, a los cuales las variables responden. Entonces podemos pensar que la no-estacionariedad de una variable corresponde, o se relaciona, con la no-estacionariedad de otra variable, por lo que existe una combinación lineal de ellas que es estacionaria. Otra manera de expresar lo anterior es que cuando dos o más variables tienen una tendencia estocástica, ellas muestran una tendencia a moverse juntas en el largo plazo.

Una ventaja de los modelos VAR, es que el principio de cointegración es invariante a extensiones del conjunto de datos, es decir que pueden incorporar nuevas series, a diferencia del análisis de regresión donde una nueva variable puede alterar las estimaciones existentes dramáticamente, además de que los modelos VAR no requieren de especificar la diferencia entre variables endógenas y exógenas.

2.2.3 Modelos MCE

Antes de que se introdujera el concepto de proceso cointegrado en la literatura econométrica existía un concepto relacionado, el concepto de modelo de corrección de errores o modelo de corrección de equilibrio, Juselius (2007). En un modelo de corrección de errores, los cambios en una variable dependen de las desviaciones a partir de una relación de equilibrio. Para ejemplificar considere que X_{1t} representa el precio de un bien en un mercado en particular y X_{2t} es el correspondiente precio del mismo bien en otro mercado. Asumase que la relación de equilibrio de las dos variables está dada por $X_{1t} = \beta_1 X_{2t}$ y que los cambios de X_{1t} dependen de las desviaciones de este equilibrio en el periodo $t-1$, es decir:

$$\Delta X_{1t} = \alpha_1(X_{1,t-1} - \beta_1 X_{2,t-1}) + \epsilon_{1t}$$

Una relación similar puede ser válida para X_{2t}

$$\Delta X_{2t} = \alpha_2(X_{1,t-1} - \beta_1 X_{2,t-1}) + \epsilon_{2t}$$

En un modelo de corrección de errores más general, los ΔX_{it} deben de depender además de cambios previos en ambas variables, como por ejemplo el siguiente modelo:

$$\Delta X_{1t} = \alpha_1(X_{1,t-1} - \beta_1 X_{2,t-1}) + \gamma_{11,1}\Delta X_{1,t-1} + \gamma_{12,1}\Delta X_{2,t-1} + \epsilon_{1t}$$
$$\Delta X_{2t} = \alpha_2(X_{1,t-1} - \beta_1 X_{2,t-1}) + \gamma_{21,1}\Delta X_{1,t-1} + \gamma_{22,1}\Delta X_{2,t-1} + \epsilon_{2t}$$

Para notar la relación entre el modelo de corrección de errores y el concepto de cointegración, supongamos que X_{1t} y X_{2t} son ambas $I(1)$. En este caso los términos en la expresión anterior que incluyen ΔX_{it} son estables, más aún, ϵ_{1t} y ϵ_{2t} son ruido blanco por lo cual también son estables, entonces la siguiente expresión es estable:

$$\alpha_i(X_{1,t-1} - \beta_1 X_{2,t-1}) = \Delta X_{it} - \gamma_{i1,1}\Delta X_{1,t-1} - \gamma_{i2,1}\Delta X_{2,t-1} - \epsilon_{it} \tag{2.4}$$

Si $\alpha_i \neq 0$, $X_{1t} - \beta_1 y_{2t}$ es estable y representa una relación de cointegración. En notación matricial el modelo anterior puede ser escrito como:

$$\Delta X_t = \alpha\beta' X_{t-1} + \Gamma_1 \Delta X_{t-1} + \epsilon_t$$

Donde $\boldsymbol{X_t}= (X_{1t}, X_{2t})'$, $\boldsymbol{\epsilon_t}= (\epsilon_{1t}, \epsilon_{2t})'$,

$$\alpha = \begin{pmatrix}\alpha_1 \\ \alpha_2\end{pmatrix}, \beta' = (1, -\beta_1), \text{ y } \Gamma_1 = \begin{pmatrix}\gamma_{11,1} & \gamma_{12,1} \\ \gamma_{21,1} & \gamma_{22,1}\end{pmatrix}$$

Reordenando 2.4 se tiene la representación $VAR(2)$:

$$X_t = (I_k + \Gamma_1 + \alpha\beta')X_{t-1} - \Gamma_1 X_{t-2} + \epsilon_t$$

Donde de nuevo las variables cointegradas son generadas por un proceso VAR. Para el caso general del proceso $VAR(p)$, k-dimensional donde sus componentes son $I(1)$ o $I(0)$:

$$X_t = \Phi_1 X_{t-1} + \cdots + \Phi_p X_{t-p} + \epsilon_t$$

Es llamado cointegrado de rango r si :

$$\Pi = -(I_k - \Phi_1 - \ldots - \Phi_p)$$

Tiene rango r, entonces Π puede factorizarse de la forma $\alpha\beta'$ con α y β siendo de dimensión $k \times r$ y de rango r. La matriz β es llamada matriz de cointegración y α es llamada la matriz de cargas, *loadings*. Si $r = 0$ entonces ΔX_t es estable y tiene una representación $VAR(p-1)$ y si $r = k$ entonces el proceso VAR no tiene raíces unitarias y es un VAR estable.

2.3 Reducción de dimensionalidad y componentes principales supervisados

La regresión por PLS, que puede ser vista como un método de componentes principales supervisados, es particularmente útil cuando el número de variables es mayor al número de observaciones. Aunque la situación descrita anteriormente no se aborda en el presente trabajo, la regresión PLS será de utilidad para realizar la estimación del pronóstico conjunto en la metodología VAR-PLS.

También la regresión PLS puede ser considerada como un método de reducción de dimensión pues crea variables que son combinaciones lineales de la variables originales -llamadas componentes- presentes en un conjunto de datos y la elección de un menor número de estas componentes implica una reducción en la dimensión del conjunto de datos con el cual se trabaja. El trabajar con un menor número de variables manteniendo fijo el tamaño de la muestra tiende a mejorar la calidad de la inferencia estadística, al igual que puede mejorar el desempeño de la regresión y en el contexto de este trabajo los pronósticos conjuntos.

La reducción de dimensionalidad no debe de confundirse con los métodos de selección de variables. Los primeros producen nuevas variables que son combinación de las demás ya sean lineales como en el caso del análisis de componentes principales (PCA) o no lineales como ISOMAP, Hastie et al. (2009); mientras que los segundos realizan una búsqueda en el conjunto potencia generado por el conjunto de variables de un conjunto de datos, por ejemplo la regresión LASSO o la búsqueda exhaustiva de subconjuntos, Hastie et al. (2009).

2.3.1 PCA

Las componentes principales son un conjunto de datos en $\Re^p$ que proporcionan una secuencia de la mejor aproximación lineal de un conjunto de datos, todas de rango $q \leq p$. Consideremos las observaciones $x_1, x_2, \ldots, x_N$ y un modelo lineal de rango q para representarlas:

$$f(\lambda) = \mu + V_q \lambda$$

Donde μ es un vector de localización y $\boldsymbol{V}_q$ es una matriz de dimensiones $p \times q$ con q vectores unitarios como columnas y λ es un vector de q parámetros. Notemos que esta es la representación paramétrica de un plano afín de rango q. Ajustar tal modelo a un conjunto de datos por medio de mínimos cuadrados implica minimizar el *error de reconstrucción*:

$$\min_{\mu, \{\lambda_i\}, V_q} \sum_{i=1}^{N} ||x_i - \mu - V_q \lambda_i||^2$$

Cuya solución está dada por:

$$\hat{\mu} = \bar{x}$$
$$\hat{\lambda_i} = V_q^t(x_i - \bar{x})$$

Lo anterior se reduce a encontrar la matriz con columnas ortogonales $\boldsymbol{V}_q$:

$$\min_{V_q} \sum_{i=1}^{N} ||(x_i - \bar{x}) - V_q V_q^t (x_i - \bar{x})||^2$$

Se puede asumir que la media muestral es cero, lo cual se logra centrando las observaciones, la matriz $\boldsymbol{H}= V_q V_q^t$ es una matriz de proyección y mapea a cada punto x_i en su reconstrucción de rango reducido, q, $\boldsymbol{H}_q x_i$, que no es más que la proyección de x_i en el subespacio generado por las columnas de $\boldsymbol{V}_q$.

Una manera de calcular la solución anterior es por medio de la descomposición en valores singulares de la matriz de datos[1] $\boldsymbol{X}$, Golub and Van-Loan (1996):

$$X = UDV^t$$

Donde los vectores columna u_i, son los vectores singulares por la izquierda y son ortogonales entre sí, de manera análoga la matriz $\boldsymbol{V}$ es ortogonal de dimensiones $p \times p$ cuyas columnas son los vectores singulares por la derecha y finalmente $\boldsymbol{D}$ es una matriz diagonal de dimensión $p \times p$ cuyos elementos no cero son los valores singulares de la matriz $\boldsymbol{X}$ tales que $d_1 \geq d_2 \geq \cdots \geq d_p \geq 0$. Para cada rango q los vectores que forman $\boldsymbol{V}_q$ consiste en las primeras q columnas de $\boldsymbol{V}$. Las columnas de $\boldsymbol{UV}$ forman las componentes principales de $\boldsymbol{X}$.

La elección óptima de un número de componentes principales es un problema abierto, existen alternativas para su elección como: el criterio del codo, el criterio de Kaiser Kaiser (1959), el *parallel analysis* propuesto en Horn (1965) y otros métodos basados en simulación Monte Carlo y resultados de matrices aleatorias, Braeken and Assen (2017).

Si bien los componentes principales tienen una práctica interpretación geométrica, poseen otras agradables propiedades como que la combinación lineal $\boldsymbol{X}v_1$ tiene la varianza más grande entre todas las combinaciones lineales de las variables originales, $\boldsymbol{X}v_2$ posee la mayor varianza con respecto a todas las combinaciones lineales que son ortogonales a la anterior,etc.

[1] A pesar de que existen diferentes algoritmos para efectuar tal descomposición de manera general la complejidad en tiempo de tal factorización es de $O(m^2 n + n^3)$ para una matriz densa de tamaño $m \times n$.

Existen otras maneras de estimar las componentes principales; por ejemplo suponiendo que poseen una distribución normal, en este contexto se les suele llamar variables latentes, se pueden estimar por medio de máxima verosimilitud y usando el método EM, Hastie et al. (2009). Con el supuesto anterior existen métodos iterativos para estimar las componentes en matrices no densas, *sparse*, Roweis (1998).

2.3.2 PLS

El método de PLS puede ser usado para regresión multivariada, así como univariada. Puede haber varias variables dependientes $Y_1, \ldots, Y_l$ y para formar la relación entre las variables $\boldsymbol{Y}$ y las variables explicativas $X_1, \ldots, X_m$, PLS construye nuevas variables, también llamadas variables latentes o componentes donde cada una de ellas es una combinación lineal de $X_1, \ldots, X_m$. Este método tiene similaridad con PCA donde las componentes principales forman las variables independientes en una regresión. La mayor diferencia entre ambos métodos radica en que las componentes de PCA son determinadas únicamente por los valores de las variables $\boldsymbol{X}$ mientras que en PLS los valores de ambos conjuntos de variables $\boldsymbol{X}$ y $\boldsymbol{Y}$ influyen en la construcción de las componentes.

La intención de PLS es formar componentes que capturen la mayor parte de la información en las variables $\boldsymbol{X}$ que sea útil para predecir $Y_1, \ldots, Y_l$, mientras se reduce la dimensionalidad del problema de regresión usando un número menor de componentes que el número total de variables en $\boldsymbol{X}$, Garthwaite (1994).

En lo siguiente primero se detalla la manera de estimar PLS para el caso univariado y posteriormente se da el algoritmo para el caso multivariado concluyendo con las propiedades de las componentes y algunas consideraciones importantes.

Las componentes derivadas por PLS pueden ser vistas como promedios ponderados de los predictores, Garthwaite (1994). Para el caso univariado supóngase que se tiene una muestra de tamaño n para las cuales se requiere estimar la relación lineal entre Y y $\boldsymbol{X}$. Para $i = 1, \ldots, n$ el i-ésimo reglón de la muestra es denotado por $(x_{i1}, \ldots, x_{im}, y_i)$. Y los valores observados de Y y X_j son denotados como y y x_j, con lo que se tiene $y = (y_1, \ldots, y_n)^t$ y para $j = 1, \ldots m$, $x_j = (x_{j1}, \ldots, x_{jn})^t$. Considérese la ecuación de regresión dada por:

$$\hat{Y} = \beta_0 + \beta_1 T_1 + \beta_2 T_2 + \cdots + \beta_p T_p \tag{2.5}$$

Donde cada componente T_k es una combinación lineal del conjunto de X_j y la correlación muestral para cada par de componentes es 0. Una ecuación conteniendo tantos parámetros es típicamente más flexible que una que contiene pocos parámetros, sin embargo, se tiene la desventaja de que los parámetros estimados pueden ser influenciados más fácilmente por el error aleatorio (o error de medición) de los datos. Es en este contexto es donde la regresión de PLS tiene ventaja al reducir el número de componentes T_i en 2.5 empleando menos variables que las que posee $\boldsymbol{X}$. Para lograr esto PLS

adopta el principio de considerar la relación entre Y y alguna variable específica de $\boldsymbol{X}$, las otras variables de $\boldsymbol{X}$ no son admitidas en tener influencia en la estimación de la relación directamente, pero se permite su influencia a través de las componentes T_k.

Para simplificar notación supongamos que en la primer iteración del algoritmo Y y $\boldsymbol{X}$ estan centradas, es decir que sus columnas tienen media cero y denotemoslas como $Y^{(1)}$ y $X^{(1)}$. En la estimación de PLS las componentes, T_k, son determinadas secuencialmente. La primer componente, T_1, pretende ser útil para predecir Y y es construida como combinación lineal de las columnas de $\boldsymbol{X^{(1)}}$, durante su construcción las correlaciones muestrales entre las columnas de $\boldsymbol{X^{(1)}}$ son ignoradas. Para obtener T_1, se realiza la regresión[1] de $Y^{(1)}$ contra la primer columna $X_1^{(1)}$, luego contra $X_2^{(1)}$ y así sucesivamente para cada $X_j^{(1)}$. Como $Y^{(1)}$ y $\boldsymbol{X^{(1)}}$ estan centrados, tienen medias muestrales iguales a cero por lo que las regresiones de mínimos cuadrados resultantes son:

$$\hat{Y}_j^{(1)} = b_j^{(1)} X_j^{(1)} \tag{2.6}$$

Donde $b_j^{(1)} = x_j^{(1)t} y^{(1)} / (x_j^{(1)t} x_j^{(1)})$. Dados los valores de $X_j^{(1)}$ para una iteración, las demas m ecuaciones de 2.6 proporcionan una manera de estimar $Y^{(1)}$. Para reconciliar estas estimaciones mientras se ignora las relaciones entre las columnas $V_j^{(1)}$, es posible tomar un simple promedio $\sum_{j=1}^{m} b_j^{(1)} V_j^{(1)}/m$ o de manera más general un promedio ponderado. Fijemos la componente T_1 igual al promedio ponderado siguiente:

$$T_1 = \sum_{j=1}^{m} w_j^{(1)} b_j^{(1)} V_j^{(1)} \tag{2.7}$$

Con la restricción de que $w_j^{(1)}$ tenga norma unitaria, aunque esto no es esencial. Como veremos posteriormente, multiplicar T_1 por una constante no afectara a los valores de las subsecuentes componentes ni las predicciones de Y. La ecuación 2.7 permite un abanico de posibilidades para la construcción de T_1, dependiendo de los pesos que se usen, consideraremos posteriormente dos formas particulares de estos pesos.

Como T_1 es un promedio ponderado de los predictores de $Y^{(1)}$, será un predictor útil para Y, pero como las variables de $\boldsymbol{X}$ potencialmente contienen información útil para predecir Y, la información en X_j que no está contenida en T_1 puede ser estimada por medio de regresar los residuos de X_j en T_1, los cuales son idénticos a los residuos de la regresión de $X_j^{(1)}$ contra T_1. De manera análoga la variabilidad de Y que no es explicada por T_1 puede ser estimada por medio de los residuos de la regresión de $Y^{(1)}$ en T_1. Estos residuos son denotados por $X_j^{(2)}$ para $X_j^{(1)}$ y por $Y^{(2)}$ para $Y^{(1)}$. La siguiente componente, T_2, es una combinación lineal de los $X_j^{(2)}$ que serán de utilidad para predecir $Y^{(2)}$.

[1] En este caso el superíndice indica el número de iteración en el que se encuentra el algoritmo.

El procedimiento se extiende iterativamente para obtener los componentes $T_2, \ldots, T_p$ donde cada componente es determinada a partir de los residuos de las regresiones de las componentes precedentes relacionando la variabilidad residual en Y con la información residual en las columnas de $\boldsymbol{X}$. específicamente suponga que $T_i, (i \geq 1)$ simplemente ha sido construido con las variables $Y^{(i)}$ y $X_j^{(i)}$ $(j = 1, \ldots, m)$ y sean los valores muestrales de $T_i, Y^{(i)}$ y las $X_j^{(i)}$ denotados por $t_i, y^{(i)}$ y $x_j^{(i)}$.

Para obtener T_{i+1}, primero se determinan $X_j^{(i+1)}$ y $Y^{(i+1)}$. Para $j = 1, \ldots, m$ se realiza la regresión de las columnas $X_j^{(i)}$ contra las T_i, teniendo $t_i^t x_j^{(i)}/(t_i^t t_i)$ como los coeficientes de la regresión, y $X_j^{(i+1)}$ se define como:

$$X_j^{(i+1)} = V_j^{(i)} - \{t_i^t v_j^{(i)}/(t_i^t t_i)\}T_i \tag{2.8}$$

Con valores muestrales $x_j^{(i+1)}$. De manera similar $Y^{(i+1)}$ es definida como $Y^{(i+1)} = Y^{(i)} - \{t_i^t y^{(i)}/(t_i^t t_i)\}T_i$, y sus valores muestrales, $y^{(i+1)}$ son los residuales de la regresión de $Y^{(i)}$ contra T_i.

La variabilidad residual en Y es $Y^{(i+1)}$ y la información restante en X_j está en $X_j^{(i+1)}$, así que el siguiente paso es realizar la regresión de $Y^{(i+1)}$ contra cada $X_j^{(i+1)}$. La j-ésima regresión produce a $b_j^{(i+1)} X_j^{(i+1)}$ como un predictor de $Y^{(i+1)}$ donde:

$$b_j^{(i+1)} = x_j^{(i+1)t} y^{i+1}/(x_j^{(i+1)t} x_j^{(i+1)}) \tag{2.9}$$

Y formando finalmente la nueva combinación lineal de los predictores:

$$T_{i+1} = \sum_j w_j^{(i+1)} b_j^{(i+1)} V_j^{(i+1)}$$

Una característica de PLS es que la correlación muestral entre cualquier par de componentes es cero, esto es consecuencia de que los residuos de una regresión estan descorrelacionadas con el regresor, así por ejemplo $X_j^{(i+1)}$ estan descorrelacionadas con T_i, además como las componentes $T_{i+1}, \ldots, T_p$ son combinación lineal de las $X_j^{(i+1)}$ entonces estan descorrelacionadas con T_i. Para completar el algoritmo el requisito de $\sum_{j=i}^k w_j^{(i)} = 1$ puede relajarse e igualar $w_j^{(i)}$ a $v_j^{(i)t} v_j^{(i)}$ para todos los pares i, j; de esta manera $w_j^{(i)} \propto Var(X_j^{(i)})$ y las componentes $T_i = \sum_j (x_j^{(i)t} y^{(i)}) X_j^{(i)} \propto \sum_j \hat{Cov}(X_j^{(i)}, Y^{(i)}) X_j^{(i)}$.

Esta es la manera usual de estimar vía PLS, la forma anterior de elegir los pesos $w_j^{(i)}$, es que sean inversamente proporcionales a las varianzas de los coeficientes $b_j^{(i)}$, también si $Var(X_j^{(i)})$ es pequeña con relación a la varianza muestral de X_j, entonces X_j es aproximadamente colineal con respecto a los componentes $T_1, \ldots, T_k$, entonces la contribución a T_i debe ser pequeña si $w_j^{(i)}$ es pequeño. La otra manera de elegir los

pesos, es que sean iguales lo que permite que cada variable X_j tenga el mismo peso para realizar predicciones. En este punto es importante destacar que la primer forma de considerar los pesos tiene como consecuencia que las componentes sean invariantes bajo rotaciones ortogonales de los datos y la segunda forma las hace invariantes bajo escala de los datos $\boldsymbol{X}$ y Y, Garthwaite (1994)

Decidir el número de componentes a incluir, p, en un modelo de regresión es un problema dificil, usualmente se puede efectuar por validación cruzada, Hastie et al. (2009), o bien en el caso multivariado por medio de la prueba de razón de verosimilitudes (suponiendo una distribución normal de los componentes), Höskuldsson (1988).

Pasando al caso multivariado existen dos maneras de realizar la estimación, realizar la regresión PLS para cada Y_i o bien hacerlo de manera múltiple como reseñamos a continuación.

El algoritmo básico para la regresión de PLS multivariado fue desarrollado en S. Wold C. Albano W. Dunn U. Edlund K. Esbensen P. Geladi S. Hellberg and Sjostrorn (1984). Se parte de dos matrices escaladas y centradas $\boldsymbol{X}$ de dimensiones $N \times M$ y $\boldsymbol{Y}$ con dimensiones de $N \times K$. El escalamiento corresponde a trabajar con matrices de correlación. El algoritmo es como sigue:

Resultado: $\boldsymbol{X}$-loadings,$\boldsymbol{Y}$-loadings

1. Iguale u a la primer columna de $\boldsymbol{Y}$
2. $w = \boldsymbol{X^t}u/(u^tu)$
3. Escale w para que tenga norma unitaria
4. $t = \boldsymbol{X}w$
5. $c = \boldsymbol{Y^t}t/(t^tt)$
6. Escale c para que tenga norma unitaria
7. $u = \boldsymbol{Y^t}c/(c^tc)$;
8. Si se ha logrado la convergencia pase a 9, en otro caso valla a 2
9. $\boldsymbol{X}$-loadings: $p = \boldsymbol{X^t}t/(t^tt)$
10. $\boldsymbol{Y}$-loadings: $q = \boldsymbol{Y^t}u/(u^tu)$
11. Realice la regresión de (u contra t): $b = u^tt/(t^tt)$
12. Estime las matrices residuales:
    $\boldsymbol{X} \leftarrow \boldsymbol{X} - tp^t$
    $\boldsymbol{Y} \leftarrow \boldsymbol{Y} - btc^t$

Algoritmo 1: PLS

En el algoritmo anterior la siguiente iteración comienza con las nuevas matrices $\boldsymbol{X}$ y $\boldsymbol{Y}$ como las matrices residuales de la previa. Las iteraciones continúan hasta que un criterio de paro se logra o bien $\boldsymbol{X}$ llega a ser la matriz cero.

Puede demostrarse, Höskuldsson (1988), que en cada nueva iteración los vectores $u^{(n)}$, $c^{(n)}$, $t^{(n)}$ y $w^{(n)}$ son proporcionales a multiplicar su iteración anterior por las matrices $\boldsymbol{YY^tXX^t}$, $\boldsymbol{Y^tXX^tY}$, $\boldsymbol{XX^tYY^t}$ y $\boldsymbol{X^tYY^tX}$ respectivamente. Lo anterior muestra que el algoritmo actúa de manera similar al *power method*, Golub and Van-Loan (1996), determinando el valor propio de cada matriz respectivamente. Cuando se logra la convergencia se tiene:

$$\begin{aligned} YY^tXX^tu &= au \\ Y^tXX^tYc &= ac \\ XX^tYY^tt &= at \\ X^tYY^tXw &= aw \end{aligned}$$

Así los vectores u, c, t y w son los vectores propios correspondientes al máximo valor propio. Lo anterior tienen entre sus consecuencias las siguientes: los vectores w son ortogonales entre sí, es decir $w_i^t w_j = 0$ para $i \neq j$; los vectores t_i también lo son, $t_i^t t_j = 0$ para $i \neq j$; los vectores t_i forman una base ortogonal del espacio generado por las columnas de $\boldsymbol{X}$ y los vectores w_i son ortogonales a los vectores p_j con $i < j$, Höskuldsson (1988).

Otra interpretación y derivación de las componentes de PLS es la de componentes con máxima covarianza. Considere dos componentes f y g en el espacio de $\boldsymbol{X}$ y $\boldsymbol{Y}$ respectivamente que cumplan lo siguiente:

$$\begin{aligned} f &= Xd \ , \ |d| = 1 \\ g &= Ye \ , \ |e| = 1 \end{aligned}$$

La covarianza muestral entre estas dos componentes está dada por $Cov(f, g) = f^t g/N$. Así con la notación del algoritmo 1 los vectores w y c satisfacen la siguiente maximización:

$$[Cov(t, u)]^2 = [Cov(Xw, Yc)]^2 = \text{máx}[Cov(f, g)]^2 \ , \ |d| = |e| = 1 \tag{2.10}$$

Para probar lo anterior se recurre a la descomposición SVD, Golub and Van-Loan (1996), de la matriz $\boldsymbol{X^tY}$:

$$X^tY = \sum_i a_i f_i g_i^t$$

El valor singular a_1 tiene la interpretación:

$$(a_1)^2 = \text{máx}(d^t X^t Y e)^2 \ , \ |d| = |e| = 1$$

Donde el máximo se obtiene cuando $d = f_1$ y $e = g_1$, en el algoritmo 1 se tiene que $w = f_1$ y $c = g_1$. Por la igualdad 2.10 se tiene que la covarianza anterior es igual a la covarianza entre las componentes t y u por lo que se puede interpretar que ellas son las componentes del espacio de $\boldsymbol{X}$ y $\boldsymbol{Y}$ que tienen máxima covarianza sobre todas las componentes en este espacio.

Para concluir esta sección solo se menciona la existencia de otra derivación de las componentes de PLS la cual se aborda desde el punto de vista de la regresión por mínimos cuadrados pesados, es decir se tiene la matriz $\boldsymbol{X^t V X}$, donde la matriz de pesos $\boldsymbol{V}$ se sustituye por la matriz $\boldsymbol{Y Y^t}$, Höskuldsson (1988).

2.4 Remuestreo y simulación

2.4.1 Bootstrap

2.4.2 Monte Carlo

CAPÍTULO 3

Metodología

3.1 Los datos

Como se menciona en Juselius (2007), los datos de las variables macroeconómicas tienen una fuerte dependencia temporal que sugiere una formulación basada en procesos estocásticos. Es útil distinguir entre:

- Variables estacionarias con una dependencia temporal a corto plazo.
- Variables no-estacionarias con dependencia temporal a largo plazo.

En la práctica es útil clasificar las variables que presentan un alto grado de persistencia en el tiempo con regreso a la media insignificante como no-estacionarias y variables que exhiben una tendencia significativa a regresar a la media como estacionarias.

El orden de integración de una variable no es en general una propiedad económica de la variable sino una conveniente aproximación estadística para distinguir entre variaciones a corto, mediano o largo plazo en los datos.

En el caso de la inflación, de acuerdo con Juselius (2007) pág. 19, existen argumentos a favor de considerarla una raíz unitaria con tendencia estocástica[1].

[1]Esto sugiere realizar y reportar pruebas de raíces unitarias a todas las variables a tratar.

Lista de Figuras

Lista de Tablas

Bibliografía

Braeken, J. and Assen, V. (2017). An empirical kaiser criterion. *Psychological Methods*, 22(3):450–466. 18

Brockwell, P. J. and Davis, R. A. (1986). *Time Series: Theory and Methods*. Springer-Verlag, Berlin, Heidelberg. 5, 6

Chan, N. H. (2010). *Time Series: Applications to Finance with R and S-Plus(R)*. Wiley Series in probability and Statistics. Wiley, 2nd edition. 6, 7, 8, 9

Frances, P. H. (2006). Forecasting 1 to h steps ahead using partial least squares. *Econometric Institute Report*, 47. 2

Garthwaite, P. H. (1994). An interpretation of partial least squares. *Journal of the American Statistical Association*, 89(425):122–127. 19, 22

Golub, G. H. and Van-Loan, C. F. (1996). *Matrix computations*. Johns Hopkins studies in the mathematical sciences. Johns Hopkins University Press, 3rd ed edition. 18, 23

González V., M. C. (2011). *Pronósticos: Metodología de Box-Jenkins*. Tipos Futura S.A. de C.V., CDMX, México. 1

Granger, C. W. J. (1981). Some properties of time series data and their use in econometric model specification. *Journal of Econometrics*, 16:121–10. 14

Haavelmo, T. (1944). The probability approch in econometrrics. *Econometrica*, 12(Supplement):1–118. 13, 14

Hamilton, J. D. (1994). *Time Series Analysis*. Princeton University Press, 1 edition. 8

Hastie, T., Tibshirani, R., and Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction, 2nd Edition*. Springer series in statistics. Springer. 17, 19, 22

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika*, 30(179-85). 18

Höskuldsson, A. (1988). Pls regression method. *Journal of Chemometrics*, 2:211 – 228. 22, 23, 24

Juselius, K. (2007). *The Cointegrated VAR Model: Methodology and Applications*. Advanced Texts in Econometrics. Oxford University Press, USA, 2 edition. 9, 13, 14, 15, 25

Kaiser, H. (1959). The application of electronic computers to factor analysis. *Meeting of Amer. Psychol. Ass.*, (Supplement):1–118. 18

Lütkepohl, H. (2006). *New Introduction To Multiple Time Series Analysis*. Springer. 12, 13

Roweis, S. (1998). Em algorithms for pca y spca. *Proceeding NIPS '97 Proceedings of the 1997 conference on Advances in neural information processing systems*, 10:626–632. 19

S. Wold C. Albano W. Dunn U. Edlund K. Esbensen P. Geladi S. Hellberg, E. J. W. L. and Sjostrorn, M. (1984). *Multivariate Data Analysis in Chemistry, in Chemometrics*. Mathematics and Statistics in Chemistry. B. R. Kowalski, Reidel Publishing Company. 22

Sax, C. and Eddelbuettel, D. (2018). Seasonal adjustment by X-13ARIMA-SEATS in R. *Journal of Statistical Software*, 87(11):1–17. 7

Sheskin, D. J. (2000). *Handbook of parametric and nonparametric statistical procedures*. Chapman Hall/CRC, 2nd ed edition. 13