CIMAT Centro de Investigación en Matemáticas, A.C.

MINIMOS CUADRADOS PARCIALES PARA SERIES DE TIEMPO FUNCIONALES NO ESTACIONARIAS: UN ENFOQUE PREDICTIVO

T E S I S

para obtener el grado de

Maestro en Ciencias con Especialidad en Probabilidad y Estadística

Presenta:

Roberto Vásquez Martínez

Director de Tesis:

Dra. Graciela González Farías

Autorización versión final

Guanajuato, Gto., 1 de julio de 2024

§ Datos de contacto y jurado

Datos del estudiante:
Roberto Vásquez Martínez.
CIMAT Sede Guanajuato.
Email: roberto.vasquez@cimat.mx

Datos del Asesor:
Dra. Graciela González Farías.
CIMAT Sede Monterrey
Email: farias@cimat.mx

Datos del Coasesor:
Dr. Israel Martínez Hernández.
Lancaster University
Email: i.martinezhernandez@lancaster.ac.uk

Datos del sinodal I:
Dr. José Ulises Márquez Urbina.
CIMAT Sede Monterrey
Email: ulises@cimat.mx

Datos del sinodal II:
Dr. Arturo Jaramillo Gil.
CIMAT Sede Guanajuato
Email: jagil@cimat.mx

A mi padre, Tomás Roberto Vásquez.

A mis abuelos, Roberto Vásquez y Carmela Robles.

A mis hermanos, Carolina, Andrés y Daniel.

Al resto de mi familia y amigos.

...Mathematics, rightly viewed, possesses not only truth, but supreme beauty a beauty cold and austere, like that of sculpture, without appeal to any part of our weaker nature, without the gorgeous trappings of painting or music, yet sublimely pure, and capable of a stern perfection such as only the greatest art can show. The true spirit of delight, the exaltation, the sense of being more than Man, which is the touchstone of the highest excellence, is to be found in mathematics as surely as in poetry.

(Bertrand Russell— Contemplation and Action, 1902-14", p.86, Psychology Press)

§ Agradecimientos

Agradezco profundamente a la Dra. Graciela González Farías por su inestimable apoyo como mentora. Durante los últimos años de mi licenciatura, su guía y confianza han sido fundamentales. Más que una asesora, ha sido una amiga y parte de mi familia.

Quiero expresar mi gratitud al CIMAT y a la Dra. Graciela González Farías por facilitar mi estancia académica en la Technische Universität München (TUM). Esta experiencia académica y personal fue invaluable para mi desarrollo.

Agradezco al Dr. Israel Martínez Hernández por su orientación, apoyo como mentor y amistad a lo largo de todo el proceso de esta tesis.

Además, agradezco a los miembros del comité evaluador de esta tesis, Dr. Arturo Jaramillo Gil, Dr. José Ulises Márquez Urbina y Dr. Emilien Antoine Marie Joly, por sus valiosos comentarios y sugerencias sobre este trabajo.

Finalmente, mi reconocimiento a todos los profesores del CIMAT en todas sus áreas, quienes contribuyeron significativamente a mi formación desde mis primeros años universitarios hasta mi posgrado. He aprendido de cada uno de ellos, tanto profesionalmente como en mi crecimiento personal.

Agradezco también a los ”probaamigos”, cuya compañía hizo esta experiencia tan enriquecedora a nivel personal y académico.

Especialmente, agradezco al CIMAT y al CONAHCYT por el apoyo económico proporcionado a través de la Beca Nacional de Posgrados Nivel Maestría, que cubrió el periodo desde agosto de 2022 hasta julio de 2024, y al proyecto CONAHCYT CBF2023-2024-3976, gracias al cual el trabajo de tesis presentado se extenderá para ser publicado en una revista arbitrada internacional.

§ Resumen

En este trabajo se presenta un estudio detallado del método de mínimos cuadrados parciales (PLS, por sus siglas en inglés) (Wold, 1966). Se describe en detalle la metodología propuesta por Muriel et al., 2012, la cual consiste en obtener componentes estacionarias en una serie de tiempo multivariada no estacionaria. En Muriel et al., 2012 se supone que esta serie de tiempo es cointegrada, pero en esta tesis se demuestra que dicha metodología es más flexible. Finalmente, se propone una innovadora generalización del algoritmo PLS para el caso en el que tanto la variable respuesta como la explicativa son funcionales, extendiendo así la metodología de Muriel et al., 2012 a series de tiempo funcionales.

En el caso multivariado, se realiza un análisis utilizando variables econométricas asociadas a la inflación y al Índice Nacional de Precios al Consumidor (INPC) en México. Para el caso funcional, se lleva a cabo un esquema de simulación para evaluar la metodología propuesta, comparándola con una metodología similar presentada en Seo, 2023a, que emplea Componentes Principales Funcionales.

§ Abstract

This work presents a detailed study of the Partial Least Squares (PLS) method (Wold, 1966). The methodology proposed by Muriel et al., 2012 is described in detail, which involves obtaining stationary components in a non-stationary multivariate time series. In Muriel et al., 2012, it is assumed that this time series is cointegrated, but this thesis demonstrates that the methodology is more flexible. Finally, an innovative generalization of the PLS algorithm is proposed for cases where both the response and explanatory variables are functional, thus extending the methodology of Muriel et al., 2012 to functional time series.

In the multivariate case, an analysis is carried out using econometric variables associated with inflation and the National Consumer Price Index (INPC) in Mexico. For the functional case, a simulation scheme is conducted to evaluate the proposed methodology, comparing it with a similar methodology presented in Seo, 2023a, which employs Functional Principal Components.

§ Índice general

§ Figuras y Tablas

Índice de figuras

Índice de tablas

§ Introducción

El estudio de fenómenos con estructura temporal es un tema recurrente en diversas áreas como la Economía, la Meteorología, entre otras. La capacidad de predecir es sumamente valiosa para comprender estos fenómenos y, en muchas ocasiones, tomar decisiones informadas.

El área de las matemáticas que formaliza los métodos estadísticos para estudiar estos fenómenos es el **Análisis de Series de Tiempo**. En este campo convergen distintas disciplinas matemáticas como la Teoría de la Probabilidad y el Análisis Complejo, además de aspectos computacionales para generar modelos informativos.

En particular, el problema de la predicción es desafiante, y el enfoque natural para abordarlo implica hacer supuestos matemáticos que, a su vez, se traducen en condiciones difíciles de encontrar en la realidad.

Precisamente, la hipótesis principal para construir esquemas predictivos es la **estacionaridad**. Sin embargo, muchos problemas reales presentan características no estacionarias. Esta dicotomía es la que inspira este trabajo de tesis.

En este trabajo se analizan dos vertientes principales: el estudio de series de tiempo vectoriales y de series de tiempo funcionales.

Para el caso vectorial, existen modelos que permiten generar predicciones bajo estacionaridad, como los modelos ARMA (Brockwell & Davis, 1991) y el modelo VAR, que se discutirá en este trabajo. También existen otros modelos, como los modelos espacio-estado (Hyndman et al., 2008). Se han desarrollado implementaciones computacionales para uso general, siendo una de las más populares Hyndman et al., 2023, la cual se puede estudiar en Hyndman y Khandakar, 2008.

Además, el estudio de fenómenos con una gran cantidad de variables es recurrente, por ejemplo, en el análisis de datos genéticos y en el análisis de propiedades químicas de materiales. Así, el problema de alta dimensionalidad es común, además del problema de no estacionaridad.

Por otro lado, el estudio de datos funcionales es un área en crecimiento en Estadística, debido a la recolección continua de información en la actualidad. En el caso funcional, se conocen algunas metodologías que permiten modelar y predecir bajo estacionaridad, como el modelo autorregresivo funcional (Ver Bosq, 2000, Capítulo 3) o la metodología más flexible propuesta en Aue et al., 2015.

Dado que muchos fenómenos se pueden conceptualizar como series de tiempo vectoriales o funcionales, es importante entender los conceptos matemáticos involucrados en cada contexto, en particular, las metodologías dedicadas a la predicción y cómo se puede lidiar con la no estacionaridad y el problema de alta dimensionalidad.

Esta tesis está organizada de la siguiente manera: en el Capítulo 1 se presenta un breve estudio de los conceptos más importantes que motivan y justifican los temas de los Capítulos 2 y 3. En el Capítulo 2 se detallan algunos de los conceptos, ideas y metodologías que se utilizan actualmente para el análisis de series de tiempo vectoriales. Uno de los conceptos más relevantes es el de *cointegración* (Engel & Granger, 1987), aplicado cuando la serie de tiempo multivariada es no estacionaria. Al final, se describe y ejemplifica el uso de las metodologías estudiadas con datos reales sobre la inflación en México.

En el Capítulo 3 se generalizan algunas de las ideas del Capítulo 2 y se propone una innovadora extensión de las ideas presentadas en Muriel et al., 2012. Esta extensión se basa en la generalización del algoritmo PLS (Wold, 1966) al contexto funcional, denominándolo algoritmo FPLS, y se establecen paralelismos entre esta metodología y la propuesta en Seo, 2023a.

Al final del Capítulo 3, se lleva a cabo un estudio de simulaciones para comparar el algoritmo FPLS con la propuesta de Seo, 2023a. Se proporcionan observaciones valiosas sobre la metodología propuesta y su potencial como herramienta general en la práctica, además de motivar nuevas líneas de investigación.

§ 1. Resultados preliminares

1.1. Mínimos cuadrados parciales

En la actualidad, la mayoría de los problemas que requieren análisis estadístico involucran la interacción de numerosas variables, lo que da lugar al recurrente problema de la dimensionalidad. Para abordar este desafío y mejorar el análisis de la información, se requieren metodologías eficaces de reducción de variables.

Entre las metodologías más populares para esta tarea se encuentran el Análisis de Componentes Principales (PCA) y el Análisis de Correlación Canónica (CCA), introducidos inicialmente en Hotelling, 1933 y Hotelling, 1936, respectivamente.

Sin embargo, PCA reduce la dimensionalidad de las variables explicativas sin considerar la variable respuesta Y en un modelo de regresión, mientras que CCA se ve limitado por la menor dimensión entre Y y X. Esto puede resultar restrictivo, especialmente dado que en muchos contextos de regresión, la dimensión de Y suele ser considerablemente menor que la de X.

El método de mínimos cuadrados parciales (PLS) tiene sus raíces en Wold, 1966, donde se propone como un enfoque iterativo basado en mínimos cuadrados para obtener componentes en sistemas con múltiples variables y para calcular correlaciones canónicas al estilo de Hotelling. Este método, propuesto por Wold, representa una generalización de CCA y ha evolucionado con el tiempo, dando lugar al algoritmo NIPALS (Wold, 1973), que se considera la versión generalizada de PLS en la actualidad (Mateos-Aparicio, 2011).

PLS es particularmente adecuado para el contexto de la regresión, donde surge para abordar problemas como la multicolinealidad y está más orientado hacia la predicción.

A continuación, describimos detalladamente este método basándonos en las explicaciones proporcionadas en Garthwaite, 1994 y Höskuldsson, 1988, enfocándonos en su aplicación dentro del modelo de regresión.

1.1.1. Caso simple

En el problema de regresión lineal simple, queremos modelar la relación de una variable aleatoria Y con respecto a un conjunto de variable aleatorias explicativas $X_1, \dots, X_m$. Se supone que $\mathbb{E}[Y|X_1, \dots, X_m]$ es una función lineal en $X_1, \dots, X_m$.

Sea una muestra $\{(y_i, x_{1i}, \dots, x_{mi}), i = 1, \dots, n\}$ de las variables Y y $X_1, \dots, X_m$.

Denotaremos en minúscula la realización de la variable aleatoria homónima correspondiente y en negritas al vector columna correspondiente a una muestra aleatoria de una misma variable.

La regresión por PLS toma la forma

$$Y = \beta_0 + \beta_1 T_1 + \cdots + \beta_p T_p + \epsilon,$$

donde $T_1, \dots, T_p$ son las componentes que obtiene el algoritmo.

El objetivo es modelar la relación de las variables explicativas X y la variable Y con una cantidad menor de parámetros $p < m$.

Las nuevas componentes T_i en la regresión serán combinaciones lineales de las variables originales $X_1, \dots, X_m$, teniendo en cuenta la variable respuesta Y. Esto contrasta con la regresión por PCA, en la cual no se considera la variable respuesta Y.

Además, es deseable que la correlación entre componentes sea 0, eso favorece la interpretación en el modelo.

Estas componentes T_i se determinarán iterativamente. En primer lugar, se considera la relación entre Y y solo una de las covariables X_j, aunque posteriormente se capturará su relación con las otras variables. De esta manera, PLS evita utilizar ecuaciones con muchas variables para la estimación de los componentes.

Empezamos considerando las variables centradas

$$U_1 = Y - \mathbb{E}[Y] \quad \text{y} \quad V_{1j} = X_j - \mathbb{E}\Big[X_j\Big],$$

y los vectores columna $\mathbf{u}_1 \in \mathbb{R}^n$ y $\mathbf{v}_{1j} \in \mathbb{R}^n$, correspondientes a las muestras aleatorias de esas variables, respectivamente.

Para construir la primer componente se realizan la regresión

$$\mathbf{u}_1 = \beta_{1j}\mathbf{v}_{1j} + \epsilon \quad \text{para} \quad j = 1, 2, \dots, m.$$

La primer componente será una media ponderada de cada uno de los predictores

$$T_1 = \sum_{j=1}^{m} w_{1j}\hat{\beta}_{1j}\mathbf{v}_{1j} \quad \text{con} \quad \sum_{j=1}^{m} w_{1j} = 1,$$

donde $\mathbf{v}_{1j} \in \mathbb{R}^n$ representa el vector columna que contiene las n realizaciones de la variable V_{1j} y $\hat{\beta}_{1j}$ corresponde al estimador del coeficiente de regresión β_{1j} .

La *información residual* que no contiene T_1 de cada X_j se obtendrá a partir de la regresión

$$\mathbf{v}_{1j} = \gamma T_1 + \epsilon \quad \text{para cada } j = 1, 2, \dots, m,$$

entonces

$$\mathbf{v}_{2j} := \mathbf{v}_{1j} - \hat{\gamma} T_1,$$

es la información residual y donde $\hat{\gamma}$ es el estimador del coeficiente de regresión γ.

La *variabilidad* que no abarca T_1 de Y se intenta capturar en la regresión

$$\mathbf{u}_1 = \eta T_1 + \epsilon,$$

dicha variabilidad estaría comprendida en el residual

$$\mathbf{u}_2 := \mathbf{u}_1 - \hat{\eta} T_1.$$

Construimos de forma análoga T_2 pero ahora usamos las muestras aleatorias contenidas en $\mathbf{u}_2$ y $\mathbf{v}_{2j}$ en el lugar de $\mathbf{u}_1$ y $\mathbf{v}_{1j}$, respectivamente.

Así,

$$\mathbf{v}_{i+1,j} = \mathbf{v}_{ij} - \frac{T_i^T \mathbf{v}_{ij}}{T_i^T T_i} T_i,$$

y

$$\mathbf{u}_{i+1} = \mathbf{u}_i - \frac{T_i^T \mathbf{u}_i}{T_i^T T_i} T_i.$$

Considerando los estimadores de cada regresión entre U_{i+1} y $V_{i+1,j}$

$$\hat{\beta}_{i+1,j} := \frac{\mathbf{v}_{i+1,j}^T \mathbf{u}_{i+1}}{\mathbf{v}_{i+1,j}^T \mathbf{v}_{i+1,j}},$$

por lo que

$$T_{i+1} := \sum_{j=1}^{m} \omega_{i+1,j}\hat{\beta}_{i+1,j}\mathbf{v}_{i+1,j},$$

con $\{\omega_{i+1,j}, j = 1, \dots, m\}$ pesos sobre el grado de importancia de cada variable en la construcción de la componente T_{i+1}.

Observación 1.1

Hacemos énfasis en los siguientes puntos:

- Se puede probar que las componentes $T_1, \dots, T_p$ no están correlacionadas, propiedad deseada pues representa la descomposición de la información en componentes al menos no dependientes linealmente.
- Decidir el número de componentes p es un problema, generalmente se realiza una análisis de validación cruzada para su estimación.

Un detalle a determinar son los pesos, estos se escogen generalmente como

$$\omega_{ij} = \mathbf{v}_{ij}^T \mathbf{v}_{ij} \propto \widehat{\text{Var}}(V_{ij}),$$

luego $\omega_{ij}\hat{\beta}_{ij} = \mathbf{v}_{ij}^T \mathbf{u}_i$ para obtener así

$$T_i := \sum_{j=1}^{m} (\mathbf{v}_{ij}^T \mathbf{u}_i) \cdot \mathbf{v}_{ij} \propto \sum_{j=1}^{m} \widehat{\text{Cov}}(V_{ij}, U_i)\mathbf{v}_{ij}, \tag{1.1}$$

donde la varianza y covarianza a la que se hace referencia en las últimas dos ecuaciones son claramente las versiones muestrales.

Observación 1.2

La motivación de estos pesos se debe a que son inversamente proporcionales a la varianza de los $\hat{\beta}_{ij}$. Si la varianza de V_{ij} es relativamente pequeña y observando que

$$V_{ij} \approx X_j - \sum_{k=1}^{i-1} a_k T_k,$$

entonces habría evidencia a favor de que $X_j \in \text{SPAN}\{T_1, \dots, T_{i-1}\}$.

La forma de construir las componentes T_i en (1.1) refuerza la intención de que se considera la relación lineal entre la variable Y y las covariables $X_1, \dots, X_m$, por lo que es natural utilizar estas componentes en un esquema de predicción.

1.1.2. Caso múltiple

Es de interés analizar la relación entre un conjunto de variables que por ilustración llamamos repuestas con respecto a un conjunto de variables explicativas. Este es precisamente el contexto de metodologías como la regresión multivariada múltiple o de CCA.

Aquí consideramos $Y_1, \dots, Y_l$ variables respuestas en relación con $X_1, \dots, X_m$ variables explicativas. Queremos hallar p componentes de forma que

$$Y_k = \beta_{k0} + \beta_{k1} T_1 + \cdots + \beta_{kp} T_p + \epsilon \quad \text{para } k = 1, \dots, l.$$

Las componentes son las mismas para cada variable explicativa, es decir, cada variable respuesta Y_k se busca sea explicada por el mismo conjunto de componentes $T_1, \dots, T_p$.

De manera análoga al caso simple, se considera a las variables centradas

$$R_{1k} = Y_k - \mathbb{E}[Y_k] \quad \text{y} \quad V_{1j} = X_j - \mathbb{E}\Big[X_j\Big],$$

y denotaremos las realizaciones de estas variables con letras minúsculas.

Para construir T_1, consideremos las matrices con los datos centrados $\mathbf{R}_1 = [\mathbf{r}_{11}, \dots, \mathbf{r}_{1l}] \in \mathbb{R}^{n \times l}$ y $\mathbf{V}_1 = [\mathbf{v}_{11}, \dots, \mathbf{v}_{1m}] \in \mathbb{R}^{n \times m}$, donde $\mathbf{r}_{1k} = [r_{1k,1}, \dots, r_{1k,n}]^T$ representa n observaciones de R_{1k} y $\mathbf{v}_{1j} = [v_{1j,1}, \dots, v_{1j,n}]^T$ representa n observaciones de V_{1j}.

Observación 1.3

Observamos que $\hat{\Sigma}_1 := \mathbf{V}_1^T\mathbf{R}_1$ es la matriz de covarianzas muestral entre las variables $[V_{11},\dots,V_{1m}]$ y $[R_{11},\dots,R_{1l}]$ que son las covarianzas entre $[Y_1,\dots,Y_l]$ y $[X_1,\dots,X_m]$.

Como se propone en Garthwaite, 1994, si $\hat{c}_1$ el vector propio asociado al valor propio más grande de $\mathbf{R}_1^T\mathbf{V}_1\mathbf{V}_1^T\mathbf{R}_1$ y definimos

$$\mathbf{u}_1 := \mathbf{R}_1\hat{c}_1,$$

la componente T_1 se construye a partir de $\mathbf{u}_1$ y $\{\mathbf{v}_{11},\dots,\mathbf{v}_{1m}\}$ como en el caso simple.

Precisamente, $\mathbf{u}_1$ es la versión muestral de U_1 en el caso simple, sin embargo, a diferencia del caso simple en el que $\mathbf{u}_1$ sólo era la muestra de Y centrada, en el caso multivariado es el primer componente principal de $\hat{\Sigma}_1$ en el espacio de columnas de Y. La intuición de esta elección es que el eigenvector $\hat{c}_1$ representa la dirección de máxima covarianza entre Y y X, entonces $\mathbf{u}_1$ representa el componente principal asociado a la variable Y.

Este mismo argumento se aplica de forma iterativa para determinar T_{i+1} y $\mathbf{v}_{i+1,j}$. Esto se ilustra en el Algoritmo 1.1.

Algoritmo 1.1 (*Garthwaite,1994*)

Inicialmente se tiene $\mathbf{R}_1 \in \mathbb{R}^{n\times l}$ y $\mathbf{V}_1 \in \mathbb{R}^{n\times m}$.
Para $i = 1,2,\dots,p$ con p el número de componentes que se desean para la regresión se itera:

1. Sea $\hat{\mathbf{c}}_i$ el vector propio correspondiente al valor propio mas grande de $\mathbf{R}_i^T\mathbf{V}_i\mathbf{V}_i^T\mathbf{R}_i$ y $\mathbf{u}_i := \mathbf{R}_i\hat{c}_i$.
2. Se construye T_i como en el caso simple a partir de $\mathbf{u}_i$ y las columnas de $\mathbf{V}_i$.
3. $\mathbf{v}_{i+1j}$ es el vector de residuos de la regresión entre $\mathbf{v}_{ij}$ y T_i como antes.
4. De manera similar, $\mathbf{r}_{i+1,j}$ es el residuo de la regresión entre $\mathbf{r}_{ij}$ y T_i.
5. Se define $\mathbf{R}_{i+1} = [\mathbf{r}_{i+1,1},\dots,\mathbf{r}_{i+1,l}]$, $\mathbf{V}_{i+1} = [\mathbf{v}_{i+1,1},\dots,\mathbf{v}_{i+1,m}]$ y se actualiza $i \leftarrow i+1$.

En cada paso, se encuentra la dirección más informativa que maximice la relación lineal entre las variables X y Y. Posteriormente, se obtiene información residual no captada por la componente. Se inicializa el algoritmo con dicha información residual como entrada.

Esta idea es similar a la descomposición de la varianza que hace PCA, sin embargo, aquí puede pensarse que se hace una descomposición de la dependencia lineal, captada con la covarianza, entre las variables.

Por otro lado, el método de mínimos cuadrados parciales también es discutido a profundidad en Höskuldsson, 1988 y cuyo algoritmo, adaptado a la notación empleada hasta el momento, es el siguiente.

Algoritmo 1.2 (*Höskuldsson,1988*)

Inicialmente se tiene $\mathbf{R}_1 \in \mathbb{R}^{n\times l}$ y $\mathbf{V}_1 \in \mathbb{R}^{n\times m}$.
Para $i = 1,\dots,m$ o mientras no se cumpla una condición de paro.
En primer lugar, sea $\mathbf{u}^{(1)}$ la primer columna de $\mathbf{R}_i$. Iteramos hasta convergencia
Para $k = 1,2,\dots$

1. $\mathbf{w}^{(k)} = \frac{\mathbf{V}_i^T\mathbf{u}^{(k)}}{(\mathbf{u}^{(k)})^T\mathbf{u}^{(k)}}$ y normalizamos $\mathbf{w}^{(k)}$.
2. Se define $\mathbf{t}^{(k)} = \mathbf{V}_i\mathbf{w}^{(k)}$ y
$$\mathbf{c}^{(k)} = \frac{\mathbf{R}_i^T\mathbf{t}^{(k)}}{(\mathbf{t}^{(k)})^T\mathbf{t}^{(k)}}.$$
Normalizamos $\mathbf{c}^{(k)}$.
3. Sea $\mathbf{u}^{(k+1)} = \mathbf{R}_i\mathbf{c}^{(k)}$ y actualizamos $k \leftarrow k+1$.

Al terminar este algoritmo, definimos

$$\hat{c}_i := \mathbf{c}^{(k)}, \quad \hat{w}_i := \mathbf{w}^{(k)} \quad \text{y} \quad T_i := \mathbf{t}^{(k)},$$

además $\mathbf{u}_i = \mathbf{R}_i \hat{c}_i$.
Posteriormente, consideramos

$$\mathbf{p}_i = \frac{\mathbf{V}_i^T T_i}{T_i^T T_i}, \quad \mathbf{q}_i = \frac{\mathbf{R}_i^T \mathbf{u}_i}{\mathbf{u}_i^T \mathbf{u}_i} \quad \text{y} \quad b_i := \frac{\mathbf{u}_i^T T_i}{T_i^T T_i}.$$

Definimos

$$\mathbf{V}_{i+1} := \mathbf{V}_i - T_i \mathbf{p}_i^T \quad \text{y} \quad \mathbf{R}_{i+1} := \mathbf{R}_i - b_i T_i \hat{c}_i^T,$$

y hacemos $i \leftarrow i+1$.
El criterio de paro que se usa es cuando $\mathbf{R}_i$ esta cerca de la matriz 0.

El Algoritmo 1.1 describe de forma intuitiva como se van construyendo las componentes $T_1, \ldots, T_p$, de forma que se maximice la dependencia lineal entre las variables Y y X. El Algoritmo 1.2 se enfoca en definir con precisión cada uno de los cálculos, esto resulta en una implementación computacional directa.

En Garthwaite, 1994 se prueba la equivalencia entre el Algoritmo 1.1 y 1.2. En Höskuldsson, 1988 se enuncian algunas de las propiedades más importantes del algoritmo PLS, que se resumen en la siguiente proposición.

Proposición 1.1

Bajo el Algoritmo 1.2 se cumple que:

1. *Los vectores $\hat{w}_1, \ldots, \hat{w}_m$ son mutuamente ortogonales.*
2. *Las componentes $T_1, \ldots, T_m$ son mutuamente ortogonales.*

Demostración. Ver Apéndice A. □

Así, el conjunto $\{\hat{w}_1, \ldots, \hat{w}_p\}$ es linealmente independiente. Se puede proyectar X al $\text{SPAN}\{\hat{w}_1, \ldots, \hat{w}_p\}$ obteniendo las componentes $T_1, \ldots, T_p$ con máxima covarianza respecto a Y.

Para terminar esta sección, enunciaremos un resultado sobre el algoritmo PLS que en lo sucesivo será muy importante.

Proposición 1.2

En cada iteración del algoritmo PLS se tiene que

$$(\hat{w}_i, \hat{c}_i) = \underset{d \in \mathbb{R}^m, e \in \mathbb{R}^l}{\text{argmax}} \widehat{\text{Cov}}(\mathbf{V}_i d, \mathbf{R}_i e)^2 \quad \textit{con } \|d\| = \|e\| = 1,$$

donde la covarianza en el lado derecho de la identidad anterior es la covarianza muestral.

Demostración. Ver Apéndice A □

Esta proposición le otorga al algoritmo PLS una interpretación geométrica directa. En cada paso del algoritmo, buscamos direcciones $\hat{w}_i$ y $\hat{c}_i$ que maximicen la relación lineal entre los vectores aleatorios $[V_{i1}, \ldots, V_{im}]^T$ y $[R_{i1}, \ldots, R_{il}]^T$, cuyas muestras aleatorias son las filas de las matrices $\mathbf{V}_i$ y $\mathbf{R}_i$, respectivamente. Esta interpretación será clave al estudiar el caso funcional.

1.2. Modelos VAR

En la sección anterior, se discutió una metodología que intenta capturar la relación entre las variables respuesta y explicativa, sin embargo, otro tipo dependencia es la temporal, que constituye el área de estu-

dio del Análisis de Series de Tiempo.

De manera similar a la sección anterior, estamos en un esquema multivariado y buscamos investigar, además, la dependencia temporal.

Por lo tanto, consideremos una $\{X_n, n \in \mathbb{Z}\}$ serie de tiempo con valores en $\mathbb{R}^m$. Un supuesto matemático necesario para el uso de muchos de los resultados en la literatura de Análisis de Series de Tiempo es la *estacionaridad*.

Definición 1.1

Decimos que una serie de tiempo $\{X_n, n \in \mathbb{Z}\}$ en $\mathbb{R}^m$ es estacionario débilmente si

1. Si $\mathbb{E}[X_n] = \mu$ con $\mu \in \mathbb{R}^m$ constante.
2. Si $\mathbb{E}\big[(X_n - \mu)(X_{n-h} - \mu)^T\big]$ existe y es una función que sólo depende de h.

En este caso, definimos

$$\Gamma_X(h) := \mathbb{E}\big[(X_n - \mu)(X_{n-h} - \mu)^T\big],$$

que la llamamos la función de autocovarianza del proceso X.

Este supuesto para procesos estocásticos y, en particular, para series de tiempo es fundamental, pues permite entender la covarianza entre las variables del proceso como sólo dependiente de la distancia temporal entre las variables y no del instante del tiempo en el que se encuentran.

Un proceso estacionario particular que podría entenderse como ruido sin estructura que modelar es un *ruido blanco*.

Definición 1.2

Decimos que una serie de tiempo $\{\epsilon_n, n \in \mathbb{Z}\}$ en $\mathbb{R}^m$ es un ruido blanco o proceso de innovación, si es un proceso estacionario con media 0 tal que

$$\Gamma_\epsilon(0) = \mathbb{E}\big[\epsilon_n \epsilon_n^T\big] \quad \text{para todo } t \quad \text{y} \quad \Gamma_\epsilon(h) = 0 \quad \text{para } h > 0.$$

Denotaremos por $\Sigma_\epsilon = \Gamma_\epsilon(0)$ que representa la matriz de covarianzas del proceso ϵ y denotamos al ruido blanco por $\epsilon \sim \mathrm{WN}(0, \Sigma_\epsilon)$.

Observación 1.4

Algunos autores también consideran la noción de *ruido blanco fuerte* que no es más que un conjunto de elementos aleatorios independientes e idénticamente distribuidos (Ver Bosq, 2000, Ejemplo 1.2).

El modelo más utilizado en Series de Tiempo en un esquema multivariado es el modelo autorregresivo vectorial (VAR por sus siglas en inglés). Un estudio a mayor profundidad se puede consultar en Lütkepohl, 2005.

Definición 1.3

Definimos al modelo VAR de orden p, denotado por VAR(p), a una solución estacionaria de la ecuación

$$X_n = \nu + A_1 X_{n-1} + \cdots + A_p X_{n-p} + \epsilon_n, \tag{1.2}$$

con $X_n \in \mathbb{R}^m$, $A_i \in \mathbb{R}^{m \times m}$ y ϵ un ruido blanco o proceso de innovación en $\mathbb{R}^m$.

Para el modelo VAR(1), la representación de media móvil infinita o MA, al igual que en el caso univariado (Ver Brockwell & Davis, 1991, Ejemplo 3.1.1) es

$$X_n = \mu + \sum_{k=1}^{\infty} A_1^k \epsilon_{n-k},$$

con $\mu = (I - A_1)^{-1}\nu$ la media del proceso X y A_1^k es la matriz A_1 multiplicada por si misma k veces.

Esta representación tiene sentido cuando los valores propios de A_1 tienen modulo menor o igual que 1, pues es cuando la serie en el lado derecho de la identidad anterior existe (Ver Lütkepohl, 2005, pp. 14).

Proposición 1.3

Los valores propios de A_1 son menores a 1 en valor absoluto si y sólo si

$$det(I_m - A_1 z) \neq 0 \quad para\ |z| \leq 1,$$

con $I_m \in \mathbb{R}^{m\times m}$ la matriz identidad.
A esta condición se le llama de estabilidad.

En el estudio de Series de Tiempo unidimensionales esta condición de estabilidad coincide con la noción de *causalidad* (Ver Brockwell & Davis, 1991, Definición 3.1.3).

Definición 1.4

Sea X un proceso que sigue un modelo VAR(1). Si det $\left(I - A_1 z^\star\right) = 0$ para algún $z^\star \in \mathbb{C}$ tal que $|z^\star| = 1$ decimos que el proceso X tiene una raíz unitaria.

La interpretación de una raíz unitaria es análoga a la expuesta en Hamilton, 1994 para el caso univariado, ahí se explica que un proceso con raíz unitaria tiene un comportamiento similar al de una caminata aleatoria, impredecible y con una variabilidad divergente en el tiempo.

Observación 1.5

Estas ideas se pueden trasladar al modelo VAR(p) escribiendo dicho modelo como un modelo VAR(1). Dicha transformación se puede consultar en Lütkepohl, 2005. Se puede concluir que, si bien aumentamos la dimensión del proceso original con variables dummies, el modelo VAR(1) es suficientemente rico en si mismo, permite reducir el problema de *overfitting* pues es un modelo más parsimonioso, además de que dicha estructura será la base para las ideas del análisis posterior.

1.2.1. Estimación

Debido a la última observación, consideraremos en esta sección el modelo VAR(1) estable en $\mathbb{R}^m$

$$X_n = \nu + A_1 X_{n-1} + \epsilon_t. \tag{1.3}$$

En la práctica, no conocemos A_1, por lo que, al igual que en un esquema de regresión tradicional, debemos obtener una estimación de los parámetros del modelo a partir de una muestra aleatoria del proceso X. Consideremos una muestra de datos $x_1, \ldots, x_T$, donde T representa el tamaño de la muestra o, en este caso, la longitud de la serie de tiempo multivariada.

Para estimar los parámetros del modelo, en primer lugar, se reescribe la ecuación (1.3) considerando las siguientes definiciones:

$$\begin{aligned}
\mathbf{X} &:= [x_1, \ldots, x_T] \in \mathbb{R}^{m\times T},\\
B &:= [\nu, A_1] \in \mathbb{R}^{m\times(m+1)},\\
z_n &:= \begin{bmatrix} 1 \\ x_n \end{bmatrix},\\
\mathbf{Z} &:= [z_0, \ldots, z_{T-1}] \in \mathbb{R}^{(m+1)\times T},\\
\mathbf{E} &:= (\epsilon_1, \ldots, \epsilon_T) \in \mathbb{R}^{m\times T}.
\end{aligned}$$

De esta manera, basándonos en la muestra aleatoria $\{x_i, i = 1, \ldots, T\}$ y con la condición inicial x_0, la versión muestral del modelo en (1.3) se expresa como:

$$\mathbf{X} = B\mathbf{Z} + \mathbf{E}.$$

Ahora, utilizando el operador "vec" y sus propiedades, descritas en el Apéndice A.12 de Lütkepohl, 2005, se tiene que:

$$\begin{aligned} \text{vec}(\mathbf{X}) &= \text{vec}(B\mathbf{Z}) + \text{vec}(\mathbf{E}) \\ &= (\mathbf{Z}^T \otimes I_m)\beta + \mathbf{e}, \end{aligned} \tag{1.4}$$

donde $\beta := \text{vec}(B)$, $\mathbf{e} = \text{vec}(\mathbf{E})$, y $\mathbf{Z}^T \otimes I_m$ representa el producto de Kronecker entre las matrices $\mathbf{Z}^T$ e I_m.

A partir de la identidad (1.4), en Lütkepohl, 2005 se deriva el estimador de mínimos cuadrados ordinarios (OLS, por sus siglas en inglés) de β, el cual es:

$$\hat{\beta}_{\text{OLS}} := \left((\mathbf{Z}\mathbf{Z}^T)^{-1}\mathbf{Z} \otimes I_m\right)\text{vec}(\mathbf{Y}).$$

Al igual que en un contexto de regresión clásico, es de interés entender la distribución del estimador $\hat{\beta}_{\text{OLS}}$. Esto, a su vez, permitirá comprender la distribución de los pronósticos del modelo (1.3).

La propiedad más importante subyacente al proceso X estudiado, que permite derivar la consistencia y normalidad asintótica del estimador $\hat{\beta}_{\text{OLS}}$, es la siguiente:

Definición 1.5

Decimos que ϵ es un ruido blanco estándar en $\mathbb{R}^m$ si es un ruido blanco fuerte y existe una constante C tal que

$$\mathbb{E}\left[\left|(\epsilon_n)_i(\epsilon_n)_j(\epsilon_n)_k(\epsilon_n)_l\right|\right] \leq C \quad \text{para todo } i,j,k,l = 1,\dots,m \quad \text{y} \quad n \in \mathbb{Z}.$$

En palabras simples, podemos derivar propiedades importantes para el estimador $\hat{\beta}_{\text{OLS}}$ si el proceso de innovación correspondiente a X tiene cuartos momentos finitos.

El resultado sobre las propiedades de consistencia y normalidad asintótica del estimador de mínimos cuadrados asociado al modelo (1.3) es

Proposición 1.4

Sea X un VAR(1) estable en $\mathbb{R}^m$ con proceso de innovación ϵ un ruido blanco estándar. Se cumple que

1. *Para $\hat{B}_{OLS} := \mathbf{X}\mathbf{Z}^T(\mathbf{Z}\mathbf{Z}^T)^{-1}$ el estimador de mínimos cuadrados de los coeficientes del modelo en (1.3) se cumple que*

   $$\hat{B}_{OLS} \xrightarrow{P} B,$$

   donde $\xrightarrow{P}$ representa convergencia en probabilidad.

2. *La covarianza muestral converge en probabilidad, es decir,*

   $$T^{-1}\mathbf{Z}\mathbf{Z}^T \xrightarrow{P} \Gamma,$$

   con Γ una matriz no singular y

   $$\sqrt{T}(\hat{\beta}_{OLS} - \beta) \xrightarrow{d} \xi \sim \mathcal{N}(0, \Gamma^{-1} \otimes \Sigma_\epsilon),$$

   donde $\xrightarrow{d}$ representa convergencia en distribución.

La prueba del resultado anterior se puede consultar en la Proposición 3.1 de Lütkepohl, 2005. Si bien el resultado se enuncia sobre la reparametrización (1.4) de (1.3), se pueden recuperar un resultado similar para ν y A_1, pues β no es más que su versión vectorizada.

1.2.2. Predicción

A continuación, se estudiará como realizar predicciones a través de un modelo VAR.

En general, dada una muestra aleatoria $\{x_n, n = 1, \dots, T\}$ del proceso X el problema de predicción consiste en realizar pronósticos a futuro del proceso X hasta un horizonte de tiempo h, es decir, inferir la distribución de las variables aleatorias $\hat{X}_{T+1}, \dots, \hat{X}_{T+h}$ que representan los pronósticos del proceso.

Formalmente, se considera la información disponible al tiempo T como

$$\Omega_T := \sigma(\{X_n, n \leq T\}),$$

donde la muestra aleatoria observada es un elemento de esta σ-álgebra, por lo que, en el problema de predicción a horizonte de tiempo h, se busca inferir la distribución de las variables predictoras $\hat{X}_{T+1}, \dots, \hat{X}_{T+h}$ condicionada a la información Ω_T.

Como es natural en un contexto de predicción, necesitamos definir un *función de pérdida* o criterio para el cual buscaremos las variables predictoras que minimicen dicha función. En general, debido a que suponemos un proceso con segundos momentos finitos, es natural pensar en variables predictoras que minimicen el error cuadrático medio (MSE por sus siglas en inglés), además, esa es una función de pérdida usual en otros contextos.

Definición 1.6

Definimos el error cuadrático medio del predictor $\hat{X}_{T+i}$ de X_{T+i} como

$$\mathsf{MSE}(\hat{X}_{T+i}) := \mathbb{E}\big[(X_{T+i} - \hat{X}_{T+i})(X_{T+i} - \hat{X}_{T+i})^T\big] \in \mathbb{R}^{m \times m}.$$

Distinguiremos dos tipos de predictores: predictor *puntual* y por *intervalo*. En un contexto de predicción, es mucho más informativo, además de conocer la media de cada una de las variables predictoras $\hat{X}_{T+1}, \dots, \hat{X}_{T+h}$, conocer también el grado de incertidumbre o un intervalo de confianza sobre los pronósticos.

Comenzamos definiendo

$$\mathbb{E}_T[X_{T+h}] := \mathbb{E}[X_{T+h} \,|\, \Omega_T].$$

Se puede probar que, para cualquier $\hat{X}_{T+i}$ predictor Ω_T-medible, se cumple que

$$\mathsf{MSE}(\hat{X}_{T+i}) \geq \mathsf{MSE}(\mathbb{E}_T[X_{T+i}]) \quad \text{para todo } i = 1, \dots, h,$$

cuya desigualdad entre matrices significa que $\mathsf{MSE}(\hat{X}_{T+i}) - \mathsf{MSE}(\mathbb{E}_T[X_{T+i}])$ es una matriz positiva.

La prueba de este hecho o la idea de como probarlo se puede consultar, por ejemplo, en el Teorema 6.3.1 de Christensen et al., 2011.

Así, el predictor óptimo de X bajo un modelo VAR(1) estable con un proceso de innovación estándar es

$$\mathbb{E}_T[X_{T+i}] = \nu + A_1 \mathbb{E}_T[X_{T+i-1}] \quad \text{para } i = 1, \dots, h. \tag{1.5}$$

A través de un argumento iterativo se tiene que

$$\mathbb{E}_T[X_{T+i}] = \nu + A_1^i x_T \quad \text{para } i = 1, \dots, h.$$

Observación 1.6

El estimador en (1.5) esta sujeto a que ϵ sea en efecto un ruido blanco estándar, en particular, que ϵ sea un proceso de variables aleatorias independientes de otra forma $\mathbb{E}_T[\epsilon_{T+i}]$ no necesariamente es 0.

Sin embargo, cuando ϵ no es un ruido blanco fuerte, se necesita proponer hipótesis adicionales para encontrar el predictor óptimo.

Sin hacer supuestos adicionales, podemos hallar el mejor predictor, bajo la función de pérdida MSE, entre el conjunto de *predictores lineales*, es decir, funciones lineales de la muestra aleatoria $\{X_n, n \leq T\}$ y ya no sobre el conjunto de funciones Ω_T-medibles que es un conjunto mucho más diverso.

Supongamos que X es la solución estacionaria del modelo VAR estable en (1.3). Notamos que para $i = 1, 2 \dots, h$

$$X_{T+i} = A_1^i X_T + \sum_{k=0}^{i-1} A_1^k \epsilon_{T+i-k}.$$

Ahora, consideremos un predictor de X_{T+i} de la forma

$$\hat{X}_{T+i} = \sum_{k=0}^{\infty} B_k x_{T-k}.$$

El objetivo es minimizar el error cuadrático medio de este predictor, utilizando precisamente que ϵ es un ruido blanco, luego ϵ_{T+i-k} es una variable no correlacionada con $\{X_n, n \leq T\}$ para todo $i = 1, \dots, h$ y $k = 1 \dots, i-1$ tenemos que

$$\text{MSE}(\hat{X}_{T+i}) = \text{Var}\left(\sum_{k=0}^{i-1} A_1^k \epsilon_{T+i-k}\right) + \text{Var}\left((A_1^i - B_0)x_T - \sum_{k=1}^{\infty} B_k x_{T-k}\right), \tag{1.6}$$

que se minimiza con $B_0 = A_1^i$ y $B_k = 0$ para $k \geq 1$.

Así, el predictor lineal óptimo es

$$\hat{X}_{T+i} = A_1^i x_T \quad \text{para} \quad i = 1, \dots, h.$$

Observación 1.7

Precisamente porque $\hat{X}_{T+i}$ es lineal, en el calculo de $\text{MSE}(\hat{X}_{T+i})$, se puede separar en los dos sumandos de lado derecho de (1.6) y sólo es necesario que ϵ sea un ruido blanco.

Observamos que el proceso de residuales utilizando el mejor predictor lineal es

$$X_{T+i} - \hat{X}_{T+i} = \sum_{k=0}^{i-1} A_1^k \epsilon_{T+i-k},$$

además, se puede deducir que

$$\text{MSE}(\hat{X}_{T+i}) = \sum_{k=0}^{i-1} A_1^k \Sigma_\epsilon (A_1^k)^T.$$

Así, observamos que a un futuro lejano

$$\text{MSE}(\hat{X}_{T+h}) \xrightarrow{h \to \infty} \Sigma_X, \tag{1.7}$$

donde $\Sigma_X := \sum_{k=0}^{\infty} A_1^k \Sigma_\epsilon (A_1^k)^T$ es la matriz de covarianzas a largo término del proceso X.

Por lo tanto, a largo plazo, el mejor predictor del proceso X es su media y su varianza es Σ_X.

Finalmente, para cuantificar el grado de incertidumbre de la predicción se puede utilizar la Proposición 1.4, aunque esto implicaría la restricción de que el proceso de innovación ϵ correspondiente a X sea un ruido blanco estándar. Bajo ese supuesto, y el hecho de que X sigue un modelo VAR(1) se deduce que la distribución de

$$(Z_{T+i})_j := \frac{(X_{T+i})_j - (\hat{X}_{T+i})_j}{\sigma_j(i)} \quad \text{para} \quad j = 1, 2, \dots, m,$$

donde $\sigma_j(i)$ es el j-ésimo elemento de la diagonal de $\text{MSE}(\hat{X}_{T+i})$, es aproximadamente normal estándar. El subíndice j representa la j-ésima coordenada del vector aleatorio correspondiente.

Cuando se supone que ϵ es un ruido blanco estándar gaussiano la distribución es exacta y este es el caso analizado en la Sección 2.2.3 de Lütkepohl, 2005.

Por lo tanto, según sea el caso, un intervalo de confianza de probabilidad de cobertura aproximada $1 - \alpha$ para la variable $(X_{T+i})_j$ es

$$C_{j,\alpha}(i) := ((\hat{X}_{T+i})_j - z_{\frac{\alpha}{2}} \sigma_j(i), (\hat{X}_{T+i})_j + z_{\frac{\alpha}{2}} \sigma_j(i)) \quad \text{para} \quad i = 1, 2, \dots, h,$$

con $z_{\frac{\alpha}{2}}$ el cuantil de probabilidad $1 - \frac{\alpha}{2}$ de una distribución normal estándar.

En la práctica se reemplaza $\sigma_j(i)$ con su estimador $\hat{\sigma}_j(i)$ obtenido a través de la estimación por OLS del modelo VAR, por la consistencia del estimador de la matriz de covarianzas descrita en la Proposición 1.4.

En esta sección se dió un resumen sobre el proceso de estimación y de predicción vía un modelo VAR, gran parte del material se encuentra en el Capítulo 2 y 3 de Lütkepohl, 2005, sin embargo, la intención aquí fue presentar de forma sucinta los resultados más importantes para nuestra investigación, realizar observaciones de valor y describir los puntos más importantes del esquema predictivo que se retomará en el transcurso de la tesis.

Adicionalmente, tanto el procedimiento de predicción como de estimación están implementados en múltiples frameworks estadísticos, por ejemplo, en la librería vars de R (Pfaff, 2008) que se utilizará en la parte computacional de este trabajo.

1.3. Elementos de Análisis Funcional

El Análisis Funcional es una disciplina matemática que se enfoca en el estudio de espacios vectoriales de dimensión infinita. Uno de los principales objetos de estudio son los espacios de funciones, como los espacios L^p, $C[0,1]$ (funciones continuas en un intervalo), y los espacios de funciones càdlàg, entre otros.

Para iniciarse en esta teoría, se puede consultar, por ejemplo, Conway, 2019; Lang, 2012; Saxe, 2002; Weidmann, 2012; Yosida, 2012.

El Análisis Funcional sirve como fundamento para un campo en crecimiento dentro de la Estadística, conocido como Análisis de Datos Funcionales (FDA por sus siglas en inglés), que constituye uno de los principales enfoques de esta investigación. FDA supone que los datos son funciones continuas, por ejemplo, perfiles de indicadores financieros.

En esta sección, nos proponemos presentar los conceptos y resultados fundamentales del Análisis Funcional que se utilizarán a lo largo de los siguientes capítulos. Comenzaremos formalizando la noción de producto interior en espacios de funciones.

Definición 1.7

Sea $\mathcal{H}$ un espacio vectorial sobre $\mathbb{C}$, decimos que $\mathcal{H}$ es un espacio de pre-Hilbert sobre $\mathbb{C}$ con producto interior $\langle\cdot,\cdot\rangle : \mathcal{H}\times\mathcal{H}\to\mathbb{C}$ si se satisface

1. $\langle\cdot,\cdot\rangle$ es una función lineal en la primer variable.
2. Para la segunda variable se cumple

$$\langle f, g_1+g_2\rangle = \langle f,g_1\rangle + \langle f,g_2\rangle \quad \text{y} \quad \langle f,\lambda g\rangle = \overline{\lambda}\langle f,g\rangle,$$

   con $\lambda\in\mathbb{C}$ y $\overline{\lambda}$ su correspondiente conjugado complejo.
3. Se cumple que

$$\langle g,f\rangle = \overline{\langle f,g\rangle}.$$

4. El producto interior es *positivo definido*, i.e

$$\langle f,f\rangle \geq 0 \quad \text{para todo } f\in\mathcal{H},$$

   y $\langle f,f\rangle = 0$ si y sólo si $f=0$.

Para un espacio de pre-Hilbert $\mathcal{H}$ se define la norma L^2 como

$$\|f\| := \sqrt{\langle f,f\rangle},$$

que en efecto cumple las propiedades de una norma (Ver Yosida, 2012, Definición 1 pp. 30). Considerando esta norma podemos definir con precisión lo que significa ser un espacio de Hilbert.

Definición 1.8

Decimos que $\mathcal{H}$ es un espacio de Hilbert en $\mathbb{C}$ si $\mathcal{H}$ es un espacio de pre-Hilbert completo bajo la norma L^2.

A continuación, presentamos un ejemplo de relevancia en Probabilidad y Estadística.

Ejemplo 1.1

Si consideramos el espacio de medida $(X, \mathcal{B}, \mu)$ se define

$$L^2(X, \mathcal{B}, \mu) = \left\{ f : X \to \mathbb{R} \text{ medible } \mid \int_X f^2 d\mu < \infty \right\},$$

que es la noción formal de espacio L^2 sobre un espacio de medida y se puede probar que este espacio de funciones es un espacio de Hilbert en $\mathbb{R}$ con producto interior

$$\langle f, g \rangle := \int_X f g \, d\mu,$$

y un caso particular de interés es cuando $(X, \mathcal{B}, \mu) = ([0,1], \mathcal{B}_{[0,1]}, \lambda)$ con $\mathcal{B}_{[0,1]}$ la σ-álgebra de Borel de $[0,1]$ y λ es la medida de Lebesgue. Este espacio particular se denotará por $L^2[0,1]$. Además, será de interés el caso en el cual $(X, \mathcal{B}, \mu) = (\Omega, \mathcal{A}, \mathbb{P})$ un espacio de probabilidad abstracto.

Si bien la teoría de Análisis Funcional esta hecha para espacios de Hilbert sobre $\mathbb{C}$, para este trabajo sólo es de interés considerar espacios de Hilbert en $\mathbb{R}$, en lo que sigue sólo hablaremos de espacios de Hilbert reales.

Así como los objetos de estudio principales en Álgebra Lineal son las transformaciones lineales en Análisis Funcional se estudian los distintos tipos de *operadores lineales*, noción que introduciremos a continuación.

Definición 1.9

Se dice que $A : \mathcal{H}_1 \to \mathcal{H}_2$ función entre dos espacios de Hilbert $\mathcal{H}_1$ y $\mathcal{H}_2$ es un operador lineal si

1. Se cumple que
$$A(f+g) = A(f) + A(g) \quad \text{para cualesquiera } f, g \in \mathcal{H}_1.$$
2. Se satisface
$$A(\lambda f) = \lambda A(f) \quad \text{para cualquier } f \in \mathcal{H}_1 \quad \text{y} \quad \lambda \in \mathbb{R}.$$

Además, se dice que es *acotado* si existe $M \geq 0$ tal que

$$\|A(f)\| \leq M \|f\| \quad \text{para todo } f \in \mathcal{H}_1.$$

Al espacio de operadores lineales acotados de $\mathcal{H}_1$ a $\mathcal{H}_2$ lo denotamos por $\mathcal{L}(\mathcal{H}_1, \mathcal{H}_2)$

Como es natural, la maldición de la dimensionalidad obliga a hacer supuestos adicionales para trabajar con los operadores lineales a como se hace en dimensión finita, estas propiedades adicionales son las que discutiremos en el resto de la sección.

Sea $\mathcal{L} := \mathcal{L}(\mathcal{H}, \mathcal{H})$ para una notación más breve. La norma de este espacio es

$$\|A\|_{\mathcal{L}} = \sup_{\|f\| \leq 1} \|A(f)\|, \quad A \in \mathcal{L}.$$

Resumimos algunas propiedades y clases de operadores en $\mathcal{L}$ de los que haremos mención a lo largo de la tesis.

Definición 1.10

Para cada $A \in \mathcal{L}$ el operador adjunto $A^\star$ es tal que

$$\langle A^\star(f), g \rangle = \langle f, A(g) \rangle \quad \text{para todo } f, g \in \mathcal{H}.$$

En lo sucesivo, será necesario que $\mathcal{H}$ sea un espacio de Hilbert separable, i.e. contenga una base ortonormal a lo más numerable. Este es un supuesto clásico en Análisis Funcional.

Definición 1.11

Se dice que $A \in \mathcal{L}$ es compacto si existe dos bases ortonormales $\{e_j, j \in \mathbb{N}\}$ y $\{f_j, j \in \mathbb{N}\}$ de $\mathcal{H}$ y una sucesión $\{\lambda_j, j \in \mathbb{N}\}$ de números reales con $\lambda_j \to 0$ tal que

$$A(g) = \sum_{j=1}^{\infty} \lambda_j \left\langle e_j, g \right\rangle f_j, \quad g \in \mathcal{H}. \tag{1.8}$$

De forma concisa

$$A = \sum_{j=1}^{\infty} \lambda_j \cdot e_j \otimes f_j.$$

El conjunto de operadores compactos lo denotamos por $\mathcal{C}$.

Observación 1.8

La descomposición en (1.8) se denomina *descomposición espectral* de A.
Ser un operador compacto es muy restrictivo, por ejemplo, el operador identidad no es compacto cuando $\mathcal{H}$ es de dimension infinita. Sin embargo, esta clase de operadores se encuentran naturalmente en el Análisis de Datos Funcionales.

Definición 1.12

Un operador compacto A se dice de *Hilbert-Schmidt* si es tal que $\sum_{j=1}^{\infty} \lambda_j^2 < \infty$. Este espacio lo denotamos por $\mathcal{S}$.
El subespacio $\mathcal{S}$ es de hecho un espacio de Hilbert con el producto interior

$$\langle S_1, S_2 \rangle_{\mathcal{S}} := \sum_{i,j} \left\langle S_1(g_i), h_j \right\rangle \cdot \left\langle S_2(g_i), h_j \right\rangle,$$

con cualesquiera $\{g_i, j \in \mathbb{N}\}$ y $\{h_j, j \in \mathbb{N}\}$ dos bases ortonormales de $\mathcal{H}$ y la norma de operador en $\mathcal{S}$ satisface

$$\|S\|_{\mathcal{S}} = \left(\sum_{j=1}^{\infty} \lambda_j^2 \right)^{\frac{1}{2}}.$$

Definición 1.13

Un operador A es simétrico o auto adjunto si

$$\langle A(f), g \rangle = \langle f, A(g) \rangle, \quad \text{para todo } f, g \in \mathcal{H},$$

y positivo si

$$\langle A(f), f \rangle \geq 0 \quad \text{para todo } f \in \mathcal{H}$$

Observamos que un operador S simétrico, positivo y de Hilbert-Schmidt tiene la siguiente descomposición

$$S(f) = \sum_{j=1}^{\infty} \lambda_j \left\langle f, e_j \right\rangle e_j,$$

con $S(e_j) = \lambda_j e_j$. Así, (λ_j, e_j) es una secuencia completa de eigenpares de S, esta observación es de suma importancia y se utilizará para realizar análisis sobre dichos operadores en un contexto estadístico, por ejemplo, esto permite generalizar la noción de PCA para el caso funcional, tema de una sección posterior.

Finalmente, introducimos el concepto de *operador nuclear*:

Definición 1.14

Un operador compacto A se dice nuclear si $\sum_{j=1}^{\infty}\left|\lambda_j\right| < \infty$ y se define la norma nuclear como

$$\|A\|_{\mathcal{N}} = \sum_{j=1}^{\infty}\left|\lambda_j\right|,$$

donde se denota por $\mathcal{N}$ al conjunto de operadores nucleares.

Observación 1.9

Se puede ver que $\mathcal{N} \subset \mathcal{S} \subset \mathcal{C} \subset \mathcal{L}$ (Ver Bosq, 2000, Identidad 1.54), lo que demuestra que en realidad nos enfocaremos en clases específicas de operadores lineales para poder generalizar algunos conceptos y resultados útiles en Estadística Multivariada, pero ahora en el Análisis de Datos Funcionales.

Otra clase de operadores de suma importancia incluso en Análisis Funcional y que tienen a su vez una gran relevancia en el Análisis de Datos Funcionales son los denominados *funcionales lineales* que definiremos a continuación.

Definición 1.15

Sea $\mathcal{H}$ un espacio de Hilbert y sea

$$\mathcal{H}^{\star} := \mathcal{L}(\mathcal{H}, \mathbb{R}),$$

el conjunto de operadores lineales acotados $l : \mathcal{H} \to \mathbb{R}$.
Al espacio $\mathcal{H}^{\star}$ lo conocemos como al espacio de dual de $\mathcal{H}$ y a sus elementos los denominamos funcionales lineales.

Para terminar esta sección enunciaremos uno de los teoremas más importantes del Análisis Funcional y que es la base de muchos de los cálculos que se hacen en el Análisis de Datos Funcionales, este es el *Teorema de Representación de Riesz*.

Teorema 1.1

Sea $\mathcal{H}$ un espacio de Hilbert y $\mathcal{H}^{\star}$ su espacio dual. Para todo $l \in \mathcal{H}^{\star}$ existe un único $f_l \in \mathcal{H}$ tal que

$$l(f) = \langle f, f_l \rangle \quad \text{para todo } f \in \mathcal{H}.$$

Los funcionales lineales serán la conexión que se utilizará para ligar la noción de elemento aleatorio en $\mathcal{H}$ con la noción clásica de variable aleatoria en $\mathbb{R}$, además, el Teorema de Representación de Riesz nos permitirá entender mejor dicha conexión.

Finalmente, un resultado de importancia para el desarrollo de modelos estadísticos en un contexto funcional es el *Teorema de la Proyección*, resultado con el que terminaremos esta sección.

Teorema 1.2

Sea $M \subset \mathcal{H}$ un subespacio cerrado del espacio de Hilbert $\mathcal{H}$ y

$$M^{\perp} := \{f \in \mathcal{H} \mid \langle f, m \rangle = 0 \quad \text{para todo } m \in M\},$$

su complemento ortogonal.
Entonces cada $f \in \mathcal{H}$ puede ser escrito de forma única como

$$f = g + h,$$

con $g \in M$ y $h \in M^{\perp}$.

De forma natural, se definen los operadores proyección a los espacios M y $M^{\perp}$ denotados por $P_M : \mathcal{H} \to M$ y $P_{M^{\perp}} : \mathcal{H} \to M^{\perp}$, respectivamente, como

$$P_M f = g \quad \text{y} \quad P_{M^{\perp}} f = h,$$

y se puede probar son operadores lineales acotados.

1.4. Probabilidad en Espacios de Hilbert

En la sección anterior introducimos todos los conceptos y elementos necesarios de Análisis Funcional para el desarrollo de la tesis.

A continuación, haremos la conexión con los conceptos clásicos de Teoría de la Probabilidad, en primera instancia, notamos que para cualquier espacio métrico, en este caso, un espacio de Hilbert $\mathcal{H}$ podemos definir una σ-álgebra de conjuntos de $\mathcal{H}$ considerando la σ-álgebra generada por la topología inducida por su norma L^2. Dicha σ-álgebra la denotamos por $\mathcal{B}_{\mathcal{H}}$ y es conocida como la σ-álgebra de Borel de $\mathcal{H}$ (Ver Billingsley, 2013, Sección 1)

Estamos listos para introducir el concepto de elemento aleatorio con valores en $\mathcal{H}$.

Definición 1.16

Sea $\mathcal{H}$ un espacio de Hilbert separable con σ-algebra de Borel $\mathcal{B}_{\mathcal{H}}$.
Una variable aleatoria $\mathcal{H}$-valuada definida en $(\Omega, \mathcal{A}, \mathbb{P})$ es un función $\mathcal{A} - \mathcal{B}_{\mathcal{H}}$-medible.

A través del Teorema de Representación de Riesz podemos obtener el siguiente criterio de medibilidad para un elemento aleatorio X en $\mathcal{H}$.

Proposición 1.5

Si $\mathcal{H}$ es un espacio de Hilbert separable, entonces $X : (\Omega, \mathcal{A}) \to (\mathcal{H}, \mathcal{B}_{\mathcal{H}})$ es una $\mathcal{H}$-valuada variable aleatoria si y solo si $l(X)$ es una variable aleatoria real para toda $l \in \mathcal{H}^{\star}$.

La conexión principal del Análisis Funcional con la Teoría de Probabilidad y en particular Estadística es el estudio de procesos de *segundo orden*, noción que definimos a continuación.

Definición 1.17

Sea $\xi = \{\xi_t, t \in [0,1]\}$ un proceso estocástico en $\mathbb{R}$. Decimos que es de segundo orden si $\mathbb{E}[\xi_t^2] < \infty$ para todo $t \in [0,1]$. Para estos procesos definimos

$$\mu(t) := \mathbb{E}[\xi_t],$$

y la función de covarianza

$$c(s,t) := \text{Cov}(\xi_s, \xi_t),$$

para $t, s \in [0,1]$.

La elección de $[0,1]$ como conjunto de índices que define al proceso estocástico es algo particular, pero lo estaremos utilizando a lo largo de la tesis, pues suele ser una convención en el estudio de procesos estocásticos a tiempo continuo y Análisis de Datos Funcionales, como se ilustra en el siguiente ejemplo.

Además, la noción de función de autocovarianza si bien esta definida a partir de un proceso estocástico en realidad también hace referencia a una función $c : [0,1]^2 \to \mathbb{R}$ tal que es simétrica y positiva definida (Ver Bosq, 2000, Ejemplo 1.4).

Ejemplo 1.2

Un proceso de segundo orden es el movimiento browniano estándar $B = \{B_t, t \in [0,1]\}$, además el moviento browniano es un proceso gaussiano (Ver Bosq, 2000, Ejemplo 1.5), propiedad que mencionamos sobre el ruido blanco de un modelo VAR que permite construir intervalos de predicción exactos.
Además, si B es un movimiento browniano sobre un espacio de probabilidad $(\Omega, \mathcal{A}, \mathbb{P})$ entonces para cada ω se tiene que la trayectoria asociada a B es un elemento de $L^2[0,1]$. Entonces, podemos entender un proceso estocástico de segundo orden como una $L^2[0,1]$-variable.

Por otro lado, presentamos dos resultados importantes que son la justificación teórica de metodologías en el Análisis Funcional de Datos como Análisis de Componentes Principales Funcionales (FPCA por sus

siglas en inglés) o Análisis de Correlación Canónica Funcional (FCCA por sus siglas en inglés) (Ver Horváth & Kokoszka, 2012, Capítulo 3 y 4).

Lema 1.1

Sea c una función de covarianza continua en $[0,1]^2$. Entonces existe una sucesión $\{\varphi_j, j \in \mathbb{N}\}$ de funciones continuas y una sucesión decreciente $\{\lambda_j, j \in \mathbb{N}\}$ de números positivos tal que

1. $\int_0^1 c(s,t)\varphi_j(s)ds = \lambda_n \varphi_n(t)$
2. $\int_0^1 \varphi_j(s)\varphi_i(s)ds = \delta_{j,i}$.

Además,

$$c(s,t) = \sum_{j=1}^{\infty} \lambda_j \varphi_j(s)\varphi_j(t),$$

donde las series convergen uniformemente en $[0,1]^2$ así

$$\sum_{j=1}^{\infty} \lambda_j = \int_0^1 c(t,t)dt < \infty.$$

Este lema, llamado *Lema de Mercer*, se puede entender como una versión funcional de la descomposición espectral de la función de covarianzas c, lo cual naturalmente implica una descomposición para el proceso estocástico cuya función de covarianza es c. Dicha descomposición se conoce como *Descomposición de Karhunen-Loeve*, resultado que se presenta a continuación.

Teorema 1.3

Sea $\xi = (\xi_t, t \in [0,1])$ un proceso estocástico de segundo orden definido en el espacio de probabilidad $(\Omega, \mathcal{A}, \mathbb{P})$ con media cero y función de covarianza continua c.
Entonces

$$\xi_t = \sum_{j=0}^{\infty} \eta_j \varphi_j(t), \quad t \in [0,1], \tag{1.9}$$

donde $\{\eta_n, n \in \mathbb{N}\}$ es una sucesión de variable aleatorias con media 0 tal que

$$\mathbb{E}\left[\eta_j \eta_i\right] = \lambda_j \delta_{j,i},$$

y donde (λ_j, φ_j) están definidas como en el Lema de Mercer. La serie en (1.9) converge uniformemente en la norma de $L^2(\Omega, \mathcal{A}, \mathbb{P})$.
Además, como se observo para el movimiento browniano , ξ define una $L^2([0,1], \mathcal{B}_{[0,1]}, \lambda)$-variable.

1.4.1. Operadores de probabilidad en espacios de Hilbert

A continuación, resumiremos algunos de los conceptos, análogos para variables aleatorias, para $\mathcal{H}$-variables.

En la sección anterior, en particular, en la Descomposición de Karhunen-Loeve se resaltó la importancia del espacio de Hilbert $L^2([0,1], \mathcal{B}_{[0,1]}, \lambda)$, por lo que en lo siguiente se puede pensar $\mathcal{H} = L^2([0,1], \mathcal{B}_{[0,1]}, \lambda)$.

Anteriormente, pensamos en un proceso ξ de segundo orden sobre un espacio de probabilidad $(\Omega, \mathcal{A}, \mathbb{P})$ definía una $L^2([0,1], \mathcal{B}_{[0,1]}, \lambda)$-variable, es decir, para cada $\omega \in \Omega$ se tiene que $\xi(\cdot, \omega) : [0,1] \to \mathbb{R}$ es un elemento de $L^2([0,1], \mathcal{B}_{[0,1]}, \lambda)$, formalizando esta idea para $\mathcal{H}$-variables en general definimos

$$L^2_{\mathcal{H}}(\Omega, \mathcal{A}, \mathbb{P}) := \left\{\xi : (\Omega, \mathcal{A}) \to (\mathcal{H}, \mathcal{B}_{\mathcal{H}}) \quad \text{medible} \;\middle|\; \mathbb{E}\left[\|\xi\|^2\right] < \infty\right\},$$

que resulta ser también un espacio de Hilbert, con producto escalar

$$(X, Y) := \mathbb{E}[\langle X, Y\rangle] \quad \text{con } X, Y \in L^2_{\mathcal{H}}(\Omega, \mathcal{A}, \mathbb{P}),$$

y en lo siguiente se abreviará $L^2_{\mathcal{H}} := L^2_{\mathcal{H}}(\Omega, \mathcal{A}, \mathbb{P})$, donde se sobreentenderá el espacio de probabilidad a menos que se especifique lo contrario.

Como nuestro objetivo final es la predicción debemos definir algunos de los elementos importantes que aparecen en el caso de dimensión finita como la esperanza condicional, la media y la varianza.

Definición 1.18

Decimos que una $\mathcal{H}$-variable X es *integrable débilmente* si $\langle X, x\rangle$ es una variable aleatoria integrable para todo $x \in \mathcal{H}$ en cuyo caso se cumple

$$\mathbb{E}[\langle X, x\rangle] = \langle x, \mathbb{E}[X]\rangle, \tag{1.10}$$

donde $\mathbb{E}[X] \in \mathcal{H}$ se denomina la esperanza débil de X.
Se dice que X es integrable o *fuertemente integrable* si $\mathbb{E}[\|X\|] < \infty$.

Observación 1.10

En la definición anterior $\mathbb{E}[X]$ está bien definido debido a que $\mathbb{E}[\langle X, \cdot\rangle] : \mathcal{H} \to \mathbb{R}$ es un funcional lineal y por el Teorema de Representación de Riesz existe un único $\mathbb{E}[X] \in \mathcal{H}$, denotado de forma conveniente, tal que se cumple (1.10).

La definición anterior formaliza la idea de media para una $\mathcal{H}$-variable, ahora formalizaremos lo que significa la esperanza condicional.

Definición 1.19

Sea $\mathcal{A}_0 \subset \mathcal{A}$ una sub-σ-álgebra de $\mathcal{A}$. La *esperanza condicional* con una $\mathcal{H}$-variable X condicionada a $\mathcal{A}_0$ es el mapeo $\mathbb{E}_{\mathcal{A}_0} : L^2_{\mathcal{H}}(\Omega, \mathcal{A}, \mathbb{P}) \to L^2_{\mathcal{H}}(\Omega, \mathcal{A}_0, \mathbb{P})$ tal que

$$\int_A \mathbb{E}_{\mathcal{A}_0}[X] d\mathbb{P} = \int_A X \, d\mathbb{P}, \quad \text{para todo } A \in \mathcal{A}_0 \text{ y } X \in L^2_{\mathcal{H}}(\Omega, \mathcal{A}, \mathbb{P}).$$

Por supuesto, la noción de esperanza condicional es válida para X integrable fuertemente, sin embargo, al considerar $X \in L^2_{\mathcal{H}}(\Omega, \mathcal{A}, \mathbb{P})$ a pesar de ser una condición más restrictiva brinda mayor sentido geométrico a la esperanza condicional y se sabe que, en este caso, se puede entender a la esperanza condicional como la proyección de X al subespacio $L^2_{\mathcal{H}}(\Omega, \mathcal{A}_0, \mathbb{P})$, que como vimos antes se puede entender como el mejor predictor de X dada la información $\mathcal{A}_0$ bajo la norma en $L^2_{\mathcal{H}}(\Omega, \mathcal{A}, \mathbb{P})$, así

$$(X, Z) = (\mathbb{E}_{\mathcal{A}_0}[X], Z) \quad \text{para todo} \quad Z \in L^2_{\mathcal{H}}(\Omega, \mathcal{A}_0, \mathbb{P}).$$

Por otro lado, introduciremos los conceptos de varianza y covarianza para variables en $L^2_{\mathcal{H}}$, estas definiciones serán de suma relevancia para el desarrollo de la tesis, en particular, para el Capítulo 3.

Definición 1.20

Sea $X \in L^2_{\mathcal{H}}$ con $\mathbb{E}[X] = 0$. Definimos $C_X : \mathcal{H} \to \mathcal{H}$ como

$$C_X(x) := \mathbb{E}[\langle X, x\rangle X],$$

el cual es un operador lineal simétrico y positivo cuya norma de operador satisface

$$\|C_X\|_{\mathcal{L}} = \sup_{\|x\| \le 1} \langle C_X(x), x\rangle, \quad x \in \mathcal{H}.$$

A C_X se le conoce como el operador de covarianza de X.

Además, se puede probar que C_X es un operador compacto por lo que su descomposición espectral se reduce a

$$C_X(x) = \sum_{j=1}^{\infty} \lambda_j \langle x, v_j\rangle v_j,$$

donde $\lambda_j = \left\langle C_X(v_j), v_j \right\rangle = \mathbb{E}\left[\left\langle X, v_j \right\rangle^2\right]$ así

$$\sum_{j=1}^{\infty} \lambda_j = \mathbb{E}\left[\|X\|^2\right] < \infty,$$

además, C_X es nuclear.

Al igual que la función de covarianza de un proceso de segundo orden, aunque el operador de covarianza se deriva de una $\mathcal{H}$-variable, en realidad este tipo de operadores se caracterizan por ciertas propiedades, como se enuncia en la siguiente proposición.

Proposición 1.6

$C \in \mathcal{L}$ es un operador de covarianza si y solo si es un operador simétrico, positivo y nuclear.

El operador de covarianza C_X para $X \in L^2_{\mathcal{H}}$ se puede entender como la generalización de la varianza de una variable aleatoria real.

Finalmente, otro de los conceptos centrales que se mencionará de forma recurrente es el de covarianza, en este caso, entre dos $\mathcal{H}$-variables X y Y, dicha noción la definimos a continuación.

Definición 1.21

Sean $X \in L^2_{\mathcal{H}_X}$ y $Y \in L^2_{\mathcal{H}_Y}$. Denominamos operadores de covarianza cruzada entre X y Y al operador $C_{X,Y} : \mathcal{H}_X \to \mathcal{H}_Y$ definido por

$$C_{X,Y}(x) := \mathbb{E}[\langle x, X\rangle Y],$$

y al operador $C_{Y,X} : \mathcal{H}_Y \to \mathcal{H}_X$

$$C_{Y,X}(y) := \mathbb{E}[\langle y, Y\rangle X].$$

Si bien en la definición consideramos que X y Y son $\mathcal{H}$-variables en espacios de Hilbert distintos, es común suponer que $\mathcal{H} = \mathcal{H}_X = \mathcal{H}_Y$ y justo en este caso se puede ver que

$$C^{\star}_{X,Y} = C_{Y,X},$$

además, se puede probar son operadores nucleares (Ver Bosq, 2000, Identidad 1.62).

1.4.2. Predicción Lineal en espacios de Hilbert

El problema de aproximar linealmente Y mediante una función lineal de variables observadas $\{X_i, i \in I\}$ es central en el análisis de regresión, discutido en la Sección 1.1 con la introducción del algoritmo PLS.

En el contexto funcional, buscamos predecir $Y \in L^2_{\mathcal{H}}$ basándonos en un conjunto de variables $\{X_i, i \in I\} \subset L^2_{\mathcal{H}}$.

En el análisis de regresión clásico, sabemos que bajo ciertas condiciones el mejor predictor lineal de Y utilizando la función de pérdida MSE es su proyección en el espacio generado por las variables $\{X_i, i \in I\}$, donde I se supone como un conjunto finito. Esto se detalla más en la estimación del modelo VAR en la Sección 1.2.

En el caso funcional, es posible que I no sea finito. Sin embargo, se puede demostrar que el predictor lineal óptimo en este escenario es la proyección de Y en el subespacio cerrado más pequeño que contiene a las variables $\{X_i, i \in I\}$.

Un subespacio de interés es

$$\text{cl SPAN}(X_i, i \in I) = \text{cl}\left\{\sum_{j \in J} a_j X_j \mid J \subset I \text{ finito y } a_j \in \mathbb{R}\right\}, \tag{1.11}$$

es decir, la cerradura del conjunto de combinaciones lineales finitas de las variables $\{X_i, i \in I\}$.

Sin embargo, en el contexto de un espacio de Hilbert de dimensión infinita, las combinaciones lineales reales de las variables $\{X_i, i \in I\}$ no abarcan todos los funcionales lineales posibles, lo cual motiva la siguiente definición.

Definición 1.22

Se dice que $\mathcal{G}$ es un subespacio cerrado hermético (LCS) de $L^2_{\mathcal{H}}$ si

1. $\mathcal{G}$ es un subespacio de Hilbert de $L^2_{\mathcal{H}}$
2. Si $X \in \mathcal{G}$ y $A \in \mathcal{L}$ entonces $A(X) \in \mathcal{G}$.

Decimos que $\mathcal{G}$ tiene media 0 si solo contiene variables con media 0.

Bajo la definición anterior, el subespacio correcto que generaliza (1.11) es el siguiente

Definición 1.23

Sea $F \subset L^2_{\mathcal{H}}$. El LCS, $\mathcal{G}_F$ generado por F es el subespacio cerrado

$$\mathcal{G}_F := \text{cl}\left\{ \sum_{i=1}^{n} A_i(X_i) \;\middle|\; A_i \in \mathcal{L},\ X_i \in F,\ i = 1,2,\dots,n,\ n \geq 1 \right\}.$$

Finalmente, podemos definir el predictor lineal para variables en $L^2_{\mathcal{H}}$.

Definición 1.24

Sea $X \in L^2_{\mathcal{H}}$ con media 0 y $\mathcal{G}$ un LCS de media 0. La variable $P_{\mathcal{G}}X$ la llamamos la predicción de X dado $\mathcal{G}$ que será el mejor predictor lineal dada la familia $\mathcal{G}$.

Al igual que en el dimensión finita, bajo normalidad el mejor predictor lineal coincide con el mejor predictor global que es la esperanza condicional.

Proposición 1.7

Sea (X, Y) una variable $\mathcal{H} \times \mathcal{H}$-variable gaussiana (Ver Bosq, 2000, pp. 33) y $\mathcal{G} = \mathcal{G}_X$ la LCS generada por X. Se cumple que

$$\mathbb{E}_{\mathcal{G}_X}[Y] = P_{\mathcal{G}_X}(Y).$$

Precisamente, utilizaremos los mismos conceptos para derivar el mejor predictor lineal, tal como hicimos con el modelo VAR, pero en su equivalente funcional: las series de tiempo funcionales, las cuales se abordarán en una sección posterior.

1.5. Introducción al Análisis de Datos Funcionales

Con los conceptos de Teoría de Probabilidad en espacios de Hilbert introducidos en las secciones anteriores, podemos ahora definir con precisión algunos de los elementos más importantes en el Análisis de Datos Funcionales. En adelante, cuando escribamos $\mathcal{H}$ para referirnos a un espacio de Hilbert, asumiremos que $\mathcal{H} = L^2[0,1]$.

Un tipo de operador en $L^2[0,1]$ es el operador *kernel* (Horváth & Kokoszka, 2012), que está relacionado con métodos homónimos como kernel PCA y regresión kernel.

Definición 1.25

Definimos el operador lineal integral Ψ en $L^2[0,1]$

$$\Psi(x)(t) = \int_0^1 \psi(t,s)x(s)ds \quad \text{para } x \in L^2[0,1],$$

entonces, decimos que es un operador con kernel ψ.

Como observamos en la Sección 1.4 los operadores de covarianza de variables en $L^2_{\mathcal{H}}$ son nucleares, en particular, son operadores de Hilbert-Schmidt, propiedad deseable también para los operadores kernel.

Proposición 1.8

El operador integral Ψ con kernel ψ es de Hilbert-Schmidt si y sólo si

$$\int_0^1 \int_0^1 \psi^2(t,s)dtds < \infty,$$

en cuyo caso

$$\|\Psi\|_S^2 = \int_0^1 \int_0^1 \psi^2(t,s)dtds < \infty,$$

Además, si el kernel ψ es tal que

- $\psi(t,s) = \psi(s,t)$, es decir, es simétrico.
- $\int_0^1 \int_0^1 \psi(t,s)x(t)x(s)dtds \geq 0$,

entonces el operador Ψ es simétrico y positivo definido.

Además, si el kernel ψ es una función continua por el Lema de Mercer admite una descomposición de la forma

$$\psi(t,s) = \sum_{j=1}^{\infty} \lambda_j v_j(t) v_j(s).$$

Ahora haremos la conexión con la clase de operadores kernel y los operadores definidos en la sección anterior para una variable $X \in L^2_{\mathcal{H}}$.

Proposición 1.9

Sea $X \in L^2_{\mathcal{H}}$ con media 0 y operador de covarianza C_X. Entonces C_X es un operador integral con kernel

$$c(t,s) := \mathbb{E}[X_t X_s],$$

y llamamos a c la función de covarianza de X noción que coincide con la introducida en la Definición 1.17.

Observamos que $c(t,s) = c(s,t)$ y

$$\int_0^1 \int_0^1 c(t,s)x(t)x(s)dtds = \mathbb{E}\left[\langle x, X\rangle^2\right] \geq 0,$$

entonces C_X es simétrico y positivo.

Observación 1.11

La propiedad del operador de covarianza C_X para $X \in L^2_{\mathcal{H}}$ es sumamente útil, ya que para realizar estimaciones muestrales de C_X se utiliza su kernel c en lugar del operador en sí mismo. Analíticamente, y por tanto computacionalmente, el kernel c es mucho más fácil de estimar a partir de una muestra.

Algunas características *teóricas* de interés para una variable $X \in L^2_{\mathcal{H}}$ son:

$$\mu(t) := \mathbb{E}[X(t)], \quad c(t,s) = \mathbb{E}[(X(t) - \mu(t))(X(s) - \mu(s))] \quad \text{para } t,s \in [0,1],$$

y el operador de covarianza C_X correspondiente con kernel c.

Sea $\{X_1, \ldots, X_T\}$ una muestra aleatoria de X independiente. Los estimadores *muestrales* de estas cantidades son, en primer lugar, para la media:

$$\hat{\mu}(t) := T^{-1} \sum_{i=1}^{T} X_i(t),$$

para la función de covarianza:

$$\hat{c}(t,s) := T^{-1} \sum_{i=1}^{T} (X_i(t) - \hat{\mu}(t))(X_i(s) - \hat{\mu}(s)),$$

y para el operador de covarianza, el estimador es el operador integral $\hat{C}_X$ con kernel $\hat{c}$.

Finalmente, en la práctica utilizaremos estos estimadores, para los cuales deseamos que posean buenas propiedades. A continuación, se resumen los principales resultados relacionados con estas propiedades.

Proposición 1.10

Sea $\{X_1, \dots, X_T\}$ una muestra aleatoria de $X \in L^2_{\mathcal{H}}$ independiente. La variable X tiene media μ y función de covarianza c. Entonces se cumple que

$$\mathbb{E}[\hat{\mu}] = \mu,$$

y

$$\mathbb{E}\big[\|\hat{\mu} - \mu\|^2\big] = O(T^{-1}).$$

Además, observamos que

$$\mathbb{E}[\hat{c}(t,s)] = \frac{T}{T-1} c(t,s).$$

Para el estimador $\hat{c}$, debemos elegir una norma apropiada para medir la discrepancia entre $\hat{c}$ y c.
Por la Proposición 1.8, se tiene directamente que C es un operador de Hilbert-Schmidt con norma

$$\|C\|^2_{\mathcal{S}} = \int_0^1 \int_0^1 c^2(t,s)\,dt\,ds.$$

A partir de esto, y dado que en virtud de la proposición anterior podemos trabajar sin mayor problema con una muestra aleatoria centrada $\{X_1, \dots, X_T\}$, se obtienen los siguientes resultados sobre las propiedades de los estimadores muestrales $\hat{c}$ y $\hat{C}_X$.

Proposición 1.11

Sea $X \in L^2_{\mathcal{H}}$ con media 0 que además $\mathbb{E}\big[\|X\|^4\big] < \infty$. Sea $\{X_1, \dots, X_T\}$ una muestra aleatoria de variables independientes con distribución común la de X.
Se cumple que

$$\mathbb{E}\Big[\big\|\hat{C} - C\big\|^2_{\mathcal{S}}\Big] \leq T^{-1}\mathbb{E}\big[\|X\|^4\big] \quad \Rightarrow \quad \mathbb{E}\Big[\big\|\hat{C} - C\big\|^2_{\mathcal{S}}\Big] = O(T^{-1}),$$

que es equivalente a

$$\mathbb{E}\left[\int_0^1 \int_0^1 (\hat{c}(t,s) - c(t,s))^2\,dt\,ds\right] = O(T^{-1}),$$

por lo que el estimador $\hat{c}$ de c es consistente en media cuadrática.

Como se vio en las Proposiciones 1.10 y 1.11, las propiedades deseables para los estimadores de la función y el operador de covarianza se obtienen bajo condiciones bastante restrictivas. Sin embargo, en contextos más generales, como cuando las variables en la muestra $\{X_1, \dots, X_T\}$ no son independientes, que es precisamente el escenario de interés en el contexto de Series de Tiempo Funcionales, es necesario investigar la validez de dichas propiedades.

1.5.1. Componentes Principales Funcionales

Finalmente, una herramienta ampliamente utilizada en Estadística y Análisis de Datos en general es el Análisis de Componentes Principales (PCA). En la Sección 1.1 discutimos una alternativa a PCA, conocida como el algoritmo PLS, enfocada en proporcionar un método de reducción de dimensionalidad con capacidad predictiva en un contexto de regresión.

Siguiendo la misma línea que en las secciones anteriores, exploraremos el análogo del método PCA en un contexto funcional, que se denomina, como era de esperar, Análisis de Componentes Principales Funcionales (FPCA por sus siglas en inglés) (Horváth & Kokoszka, 2012).

Para un vector aleatorio $X \in \mathbb{R}^m$, la idea de PCA es obtener componentes que sean combinaciones lineales de las variables que lo componen, concentrando la mayor variabilidad posible.

Consideremos ahora una variable $X \in L^2_{\mathcal{H}}$, donde C_X representa su operador de covarianza. Recordemos que C_X es un operador compacto, simétrico y positivo definido con una descomposición espectral dada por:

$$C_X(x) = \sum_{j=1}^{\infty} \lambda_j \langle x, v_j \rangle v_j \quad \text{para todo } x \in \mathcal{H}.$$

Los conjuntos $\{\lambda_j, j \in \mathbb{N}\}$ y $\{v_j, j \in \mathbb{N}\} \subset \mathcal{H}$ son conocidos como los eigenvalores y eigenfunciones, respectivamente, del operador C_X.

En la práctica, estimamos los p eigenvalores más grandes y asumimos que

$$\lambda_1 > \lambda_2 > \ldots > \lambda_p > \lambda_{p+1},$$

indicando que los primeros p eigenvalores son distintos de 0.

Además, consideramos las eigenfunciones normalizadas, es decir, $\|v_j\| = 1$.

Observación 1.12

Lo anterior no determina el signo de v_j. Si $\hat{v}_j$ es un estimador esperamos que $\hat{c}_j \hat{v}_j$ es cercano a v_j donde

$$\hat{c}_j := \text{sign}\left\langle \hat{v}_j, v_j \right\rangle,$$

sin embargo, $\hat{c}_j$ no se puede estimar de los datos.
Debe asegurarse que los estadísticos a utilizar no dependan de $\hat{c}_j$.

Los estimadores muestrales de los eigenvalores y eigenfunciones de una muestra aleatoria $\{X_1, \ldots, X_T\}$ de la variable X se obtienen a partir de la descomposición espectral del operador estimador $\hat{C}_X$ y su kernel correspondiente $\hat{c}$, definido en la sección anterior. Es decir, los primeros p eigenvalores y eigenfunciones estimados satisfacen la relación:

$$\int \hat{c}(t,s)\hat{v}_i(s)\,ds = \hat{\lambda}_i \hat{v}_i(t) \quad \text{para } i = 1, 2, \ldots, p,$$

y de manera análoga al caso de dimensión finita, el lado derecho de esta identidad se denomina el i-ésimo componente principal, para $i = 1, 2, \ldots, p$.

Es crucial cumplir con las mismas condiciones discutidas en la sección anterior para garantizar la consistencia de estos estimadores, como se detalla en la siguiente proposición.

Proposición 1.12

Consideremos $X \in L^2_{\mathcal{H}}$ con media 0 y tal que $\mathbb{E}\{\|X\|^4\} < \infty$.
Supongamos que los primeros p eigenvalores de X son positivos, es decir, $\lambda_1 > \lambda_2 > \ldots > \lambda_p > \lambda_{p+1}$.
Para una muestra aleatoria $\{X_1, \ldots, X_T\}$ de variables independientes con distribución X, se cumple que para $1 \le i \le p$:

$$\limsup_{T\to\infty} T \cdot \mathbb{E}\{\|\hat{c}_i \hat{v}_i - v_i\|^2\} < \infty,$$

y

$$\limsup_{T\to\infty} T \cdot \mathbb{E}\{|\hat{\lambda}_i - \lambda_i|^2\} < \infty.$$

Esta proposición implica que

$$T^{\frac{1}{2}}(\hat{c}_i \hat{v}_i - v_i) = o_p(1),$$

y

$$T^{\frac{1}{2}}(\lambda_i - \hat{\lambda}_i) = o_p(1),$$

que es directamente la consistencia deseada en los estimadores.

De hecho, los resultados presentados en esta sección son válidos para cualquier operador simétrico de Hilbert-Schmidt en $\mathcal{H}$.

Si Ψ tiene la descomposición espectral

$$\Psi(x) = \sum_{j=1}^{\infty} \lambda_j \langle x, v_j \rangle v_j,$$

entonces

$$\langle \Psi(x), x \rangle = \langle \sum_{j=1}^{\infty} \lambda_j \langle x, v_j \rangle v_j, x \rangle = \sum_{j=1}^{\infty} \lambda_j \langle x, v_j \rangle^2 \quad \text{para todo } x \in \mathcal{H}.$$

Usando la Identidad de Parseval (Ver Yosida, 2012, Teorema 2 pp. 86), obtenemos

$$\|x\|^2 = \sum_{j=1}^{\infty} \langle x, v_j \rangle^2 = 1,$$

de donde se deduce el siguiente resultado a partir de estas dos identidades.

Teorema 1.4

Sea Ψ un operador simétrico y positivo definido de Hilbert-Schmidt con eigenfunciones v_j y eigenvalores λ_j, donde se cumple que

$$\lambda_1 > \lambda_2 > \ldots > \lambda_p > \lambda_{p+1}.$$

Entonces, para $1 \leq i < p$, se tiene que

$$\sup\{\langle \Psi(x), x \rangle \mid \|x\| = 1, \langle x, v_j \rangle = 0 \text{ para } 1 \leq j \leq i-1\} = \lambda_i,$$

y este valor se alcanza en v_i, siendo único salvo el signo.

Esta propiedad de los eigenvalores y eigenfunciones de un operador simétrico y positivo definido de Hilbert-Schmidt será fundamental para respaldar la metodología propuesta en esta tesis.

En resumen, el método FPCA implica encontrar los eigenvalores y eigenfunciones del operador C_X o del operador $\hat{C}_X$ dada una muestra aleatoria $\{X_1, \ldots, X_T\}$. Específicamente, consideramos:

- Las eigenfunciones $\hat{v}_j$ de $\hat{C}_X$ como los Componentes Principales Funcionales Empíricos (EFPC).
- Los Componentes Principales Funcionales Teóricos (FPC) son las eigenfunciones v_j de C_X.

Observación 1.13

El EFPC estima los FPC bajo las condiciones de regularidad que da la Proposición 1.12.

Finalmente, la intuición detrás del método FPCA se desprende del Teorema 1.4 aplicado al operador de covarianza C_X, y está relacionada con una descomposición de C_X, que a su vez implica una descomposición de la variabilidad.

Podemos probar que

$$\langle \hat{C}_X(x), x \rangle = T^{-1} \sum_{i=1}^{T} \langle X_i, x \rangle^2,$$

lo cual se interpreta como la variabilidad de los datos en la *dirección* de la función x.

Por otro lado, para maximizar la variabilidad explicada en la dirección x, debemos encontrar el máximo de

$$\langle C_X(x), x \rangle \quad \text{sujeto a} \quad \|x\|^2 = 1.$$

Según el Teorema 1.4, este máximo se alcanza en v_1, que corresponde al primer Componente Principal Funcional (FPC).

El razonamiento anterior es completamente análogo cuando se usa el estimador $\hat{C}_X$, en cuyo caso obtenemos el primer Componente Principal Funcional Empírico (EFPC).

Por lo tanto, tenemos la interpretación estándar de los componentes principales: son direcciones ortogonales entre sí donde la proyección maximiza la variabilidad, ya sea en el contexto teórico o estimado.

Aunque esta sección ha cubierto los conceptos teóricos fundamentales del Análisis de Datos Funcionales, con un método destacado como FPCA, se puede profundizar más consultando Horváth y Kokoszka, 2012; Ramsay et al., 2009; Ramsay y Silverman, 2005. Además, en el ámbito computacional, existen frameworks que facilitan el análisis eficiente de este tipo de datos, como la paquetería fda (Ramsay, 2023), la cual utilizaremos en las implementaciones computacionales de esta tesis.

1.5.2. Series de Tiempo Funcionales

Por otro lado, en esta última sección se introduce el objeto de estudio principal de esta tesis: las *series de tiempo funcionales.*

Hasta este punto, hemos mencionado que en el contexto clásico de datos funcionales, se tiene una muestra $\{X_1, \dots, X_T\}$ de variables en $L^2_{\mathcal{H}}$ independientes e idénticamente distribuidas.

Sin embargo, es posible considerar que el conjunto de funciones aleatorias $\{X_1, \dots, X_T\}$ está indexado de manera que corresponde a muestras aleatorias tomadas en intervalos consecutivos de tiempo, por ejemplo, días. Es decir, X_i representa una función aleatoria observada en el día i.

En Horváth y Kokoszka, 2012 se ilustran algunos casos donde los datos pueden entenderse de esta manera, como curvas diarias de datos financieros, patrones diarios geofísicos o variables ambientales.

Definición 1.26

Se dice que $\{X_n, n \in \mathbb{Z}\}$ es una serie de tiempo funcional si constituye un proceso estocástico de variables en $\mathcal{H}$, donde el índice de las variables representa el orden cronológico en unidades de tiempo.

Observación 1.14

Suponemos en lo subsecuente que $\mathbb{E}[X_n] = 0$ para todo n.

Recordamos las extensiones del concepto de varianza y covarianza en el contexto funcional.

Definición 1.27

Definimos el kernel del operador de covarianza como

$$\gamma_{n,n+h}(t,s) := \mathrm{Cov}(X_n(t), X_{n+h}(s)).$$

De esta manera, podemos definir el operador de covarianza cruzada correspondiente, el cual generaliza la función de autocovarianza de una serie de tiempo en el contexto clásico.

Como se discutió en la Sección 1.2, la estacionaridad es una propiedad fundamental en el análisis de series de tiempo funcionales.

Definición 1.28

Sea $\{X_n, n \in \mathbb{Z}\}$ una serie de tiempo funcional con valores en $\mathcal{H}$. Decimos que es estacionaria débilmente si $\{X_n, n \in \mathbb{Z}\} \subset L^2_{\mathcal{H}}$ y además

1. Si $\mathbb{E}[X_n] = \mu \in \mathcal{H}$ constante.
2. Si el operador de covarianza entre las variables X_n y X_{n+h} no depende de n.

Denotamos por $\Gamma_X(h)$ al operador de covarianza cruzada entre X_n y X_{n+h}, lo denominamos el operador de autocovarianza de la serie X.
Al kernel del operador $\Gamma_X(h)$

$$\gamma_h(t,s) := \mathrm{Cov}(X_0(t), X_h(s)),$$

lo llamamos la función de autocovarianza de la serie X.

La definición es completamente análoga a la definición de estacionaridad para una serie de tiempo en dimensión finita, adaptando los conceptos correspondientes a su versión funcional. De manera similar, tenemos la noción de ruido blanco.

Definición 1.29

Sea $\{\epsilon_n, n \in \mathbb{Z}\}$ un serie de tiempo funcional con valores en $\mathcal{H}$. Decimos que es un ruido blanco si es una serie de tiempo estacionaria con media 0 tal que

$$\Gamma_\epsilon(0) = C_{\epsilon_0} \quad \text{y} \quad \Gamma_\epsilon(h) = 0 \quad \text{para } h > 0.$$

Por otro lado, observamos que la covarianza del proceso X varía dependiendo de la distancia entre las

variables que estamos considerando. Para llevar a cabo un análisis similar al realizado para el operador de covarianza de una muestra de variables independientes e idénticamente distribuidas, debemos considerar un operador de covarianza que capture la dependencia temporal de la serie funcional X. Esto motiva la siguiente definición.

Definición 1.30

Sea $\{X_n, n \in \mathbb{Z}\}$ una serie de tiempo funcional estacionaria con valores en $L^2_{\mathcal{H}}$. El operador de *covarianza a largo término* se define como

$$\Gamma(x)(s) := \int_0^1 \gamma(t,s)x(t)dt, \quad x \in \mathcal{H},$$

donde

$$\gamma(t,s) = \sum_{h=1}^{\infty} \gamma_h(t,s).$$

Observación 1.15

La estacionaridad no garantiza automáticamente la existencia del operador Γ.
Para una serie de tiempo estacionaria con representación MA,

$$X_n = \sum_{k=0}^{\infty} A_k(\epsilon_{n-k}),$$

donde $\{\epsilon_n, n \in \mathbb{Z}\}$ es un ruido blanco fuerte en $\mathcal{L}^2_{\mathcal{H}}$, una condición suficiente adicional es que $\sum_{k=0}^{\infty} \|A_k\| < \infty$. Esto garantiza la existencia de $A := \sum_{k=0}^{\infty} A_k$, lo cual implica que Γ existe y está dado por

$$\Gamma = AC_\epsilon A^*,$$

donde C_ϵ es el operador de covarianza de cualquiera de las variables del proceso ϵ, y A^* representa el operador adjunto de A (Martínez-Hernández et al., 2022).

Para una muestra aleatoria $\{X_1, \dots, X_T\}$ de una serie de tiempo funcional, la falta de distribución idéntica e independencia entre las variables requiere inicialmente definir la noción de covarianza general para la serie. Además, debemos abordar el efecto de la dependencia temporal en las propiedades de los estimadores muestrales, como el operador de covarianza muestral a largo plazo.

Sin embargo, mientras que el operador de covarianza a largo plazo se centra en comprender el comportamiento de la serie de tiempo funcional a largo plazo, el objetivo principal de este trabajo es estudiar su comportamiento a corto plazo y desarrollar un esquema predictivo en consecuencia. La discusión sobre la covarianza a largo plazo se presenta debido a su relevancia en la investigación actual sobre series de tiempo funcionales, como se explica en Horváth y Kokoszka, 2012; Martinez Hernandez y Genton, 2019; Martínez-Hernández et al., 2022.

Para obtener resultados análogos a las Proposiciones 1.11 y 1.12 en el contexto de una serie de tiempo funcional $\{X_1, \dots, X_T\}$, es fundamental que la serie de tiempo subyacente X sea L^4-m-aproximable (Ver Horváth & Kokoszka, 2012, Definición 16.1). La intuición detrás de esto radica en la capacidad de aproximar la serie X mediante procesos $X^{(m)}$ que son m-dependientes, lo que implica que la dependencia temporal entre las variables desaparece para distancias mayores a m; es decir, la dependencia temporal no es perpetua.

Para una serie de tiempo funcional estacionaria X con media cero, definimos $C_X := \Gamma_X(0)$. Los resultados análogos a las Proposiciones 1.11 y 1.12 para el análisis espectral del operador de covarianza estimado $\hat{C}_X$, con kernel dado por

$$\hat{c}(t,s) = T^{-1} \sum_{i=1}^{T} X_i(t)X_i(s),$$

se presentan a continuación.

Proposición 1.13

Sea $\{X_n, n \in \mathbb{Z}\}$ una serie de tiempo funcional estacionaria de media 0. Supongamos que $\{X_n, n \in \mathbb{Z}\} \subset L^4_{\mathcal{H}}$ y es L^4-m-aproximable. Existe una constante U_X que no depende de T tal que

$$\mathbb{E}\left[\left\|\hat{C}_X - C_X\right\|_S\right] \leq U_X T^{-1}.$$

Proposición 1.14

Sea $\{X_n, n \in \mathbb{Z}\}$ una serie de tiempo funcional estacionaria de media 0. Supongamos $\{X_n, n \in \mathbb{Z}\} \subset L^4_{\mathcal{H}}$ y es L^4-m-aproximable.
Supongamos que los eigenvalores de C_X son tales que

$$\lambda_1 > \lambda_2 > \ldots > \lambda_p > \lambda_{p+1},$$

es decir, los primeros p eigenvalores de X son positivos.
Para una muestra aleatoria $\{X_1, \ldots, X_T\}$ de de la serie $\{X_n, n \in \mathbb{Z}\}$ se cumple para $1 \leq i \leq p$

$$\limsup_{T\to\infty} T \cdot \mathbb{E}\left[\|\hat{c}_i\hat{v}_i - v_i\|^2\right] < \infty,$$

y

$$\limsup_{T\to\infty} T \cdot \mathbb{E}\left[\left|\hat{\lambda}_i - \lambda_i\right|^2\right] < \infty.$$

Estos resultados ofrecen la posibilidad de aplicar FPCA, por ejemplo, a muestras de series de tiempo estacionarias. Sin embargo, el enfoque principal de esta tesis son las series de tiempo no estacionarias. En adelante, se explicará cómo la metodología de FPCA o la alternativa propuesta en esta tesis aún pueden aplicarse en el contexto no estacionario, bajo ciertos supuestos.

§ 2. Componentes estables de una serie de tiempo vectorial

Como vimos en la sección anterior, la estacionaridad es una propiedad imprescindible para las metodologías de predicción, ya que permite controlar el error y proporciona cierta certeza en los pronósticos.

En esta sección, analizaremos cómo abordar el problema de la no estacionaridad, un fenómeno muy común en la práctica, con el fin de desarrollar esquemas predictivos efectivos.

2.1. Introducción al concepto de cointegración

Consideremos $\{X_n, n \in \mathbb{Z}\}$ como una serie de tiempo en $\mathbb{R}^m$. Cada una de las m componentes de la serie $\{X_n, n \in \mathbb{Z}\}$ puede ser vista como una serie que interactúa con las demás. Este tipo de relaciones entre variables se estudian en diversos campos, como el análisis económico y ambiental.

Por ejemplo, en el ámbito económico, una variable económica vista como una serie de tiempo puede fluctuar independientemente, pero se puede esperar que dos series de tiempo se muevan de manera que no estén muy alejadas una de la otra. Típicamente, en la teoría económica, existen fuerzas que mantienen a las variables económicas, vistas como series, unidas. Ejemplos de este fenómeno pueden ser las tasas de interés, las inversiones de capital y el gasto. Una idea similar consiste en considerar las "relaciones de equilibrio", que pueden entenderse como las fuerzas que mantienen, en este caso, a la economía en equilibrio (Engel & Granger, 1987).

En Economía, una serie de variables económicas $\{X_n, n \in \mathbb{Z}\}$ con valores en $\mathbb{R}^m$ se dice que está en equilibrio si satisface la condición:

$$\beta^T X_n = 0 \quad \text{para todo } n,$$

donde $\beta \in \mathbb{R}^m$ es un vector dado.

Sin embargo, esta condición es bastante restrictiva y puede que en algunos momentos n no se cumpla el equilibrio. Por lo tanto, definimos el proceso:

$$Z_n := \beta^T X_n,$$

que representa el error con respecto al equilibrio.

Comúnmente, cuando se trabaja con una serie de tiempo, se realiza la diferenciación un número finito de veces para obtener un proceso estacionario, que se puede aproximar por un proceso autorregresivo.

Definición 2.1

Sea $\{X_n, n \in \mathbb{Z}\}$ una serie de tiempo con valores en $\mathbb{R}^m$.
El operador diferencia Δ se define como

$$\Delta X_n := X_n - X_{n-1} \quad \text{para todo } n \in \mathbb{Z}.$$

Se dice que la serie $\{X_n, n \in \mathbb{Z}\}$ es integrada de orden $d \in \mathbb{N}$ si d es el número natural más pequeño tal que la serie $\{\Delta^d X_n, n \in \mathbb{Z}\}$ es estacionaria. Denotamos esta propiedad como $X \sim \mathrm{I}(d)$.
Aquí, Δ^d representa la aplicación del operador diferencia d veces.

Observación 2.1

Es común considerar los casos $d = 0$ y $d = 1$ (una raíz unitaria). En el caso $d = 0$, se tiene un proceso con un comportamiento estacionario, mientras que en el caso $d = 1$, la varianza del proceso diverge y las innovaciones tienen un efecto permanente en el proceso.

En primera instancia, tener un conjunto de series $\mathrm{I}(1)$ resulta problemático, debido a que, a diferencia del caso estacionario, se presenta un comportamiento caótico en la varianza del proceso.

No obstante, al considerar la noción de equilibrio entre diferentes variables dentro de un vector de series económicas, es posible combinar estos comportamientos impredecibles para obtener un proceso estable. Esta es precisamente la idea de *cointegración* entre series de tiempo, un concepto que fue delineado con precisión en el influyente artículo de Engel y Granger, 1987.

Definición 2.2

Sea $\{X_n, n \in \mathbb{Z}\}$ una serie de tiempo en $\mathbb{R}^m$.
Consideremos

$$X_n = [X_{1,n}, \dots, X_{m,n}]^T \quad \text{para todo} \quad n \in \mathbb{Z},$$

donde $X_{i,n}$ es la coordenada i-ésima del vector X_n.
Denotemos $X_i := \{X_{i,n}, n \in \mathbb{Z}\}$ a la serie de tiempo univariada correspondiente a las i-ésimas coordenadas del proceso $\{X_n, n \in \mathbb{Z}\}$.
Se dice que las $X_1, \dots, X_m$ son series cointegradas de orden d, b, lo que denotamos por $X \sim \mathrm{CI}(d, b)$, si

1. Si $X_i \sim \mathrm{I}(d)$ para todo $i = 1, \dots, m$.
2. Existe $\beta \in \mathbb{R}^m$ con $\beta \neq 0$ tal que
$$Z_n := \beta^T X_n \sim \mathrm{I}(d - b).$$

El vector β se denomina como una relación o vector de cointegración.

Bajo la definición anterior, las relaciones de cointegración no necesariamente son únicas. Si existen r vectores de cointegración que son linealmente independientes, entonces r es el rango de cointegración.

Existe una relación estrecha entre la cointegración y los modelos de corrección del error (Ver Lütkepohl, 2005, Capítulo 6). La idea fundamental es que la proporción de desequilibrio (representada por Z_n) se corrige en el siguiente periodo $n + 1$.

Definición 2.3

Sea $\{X_n, n \in \mathbb{Z}\}$ una serie de tiempo en $\mathbb{R}^2$ y sus componentes son $\mathrm{I}(1)$ o $\mathrm{I}(0)$ únicamente.
Decimos que X tiene una representación de corrección del error si se puede expresar como

$$A(L)(1 - L)X_n = -\alpha Z_{n-1} + \epsilon_n,$$

con ϵ un proceso estacionario multivariado que representa el error, $A(0) = I_2$ y $A(1)$ de entradas finitas, $Z_n = \beta^T X_n$ y $\alpha \in \mathbb{R}^2$.
Aquí L es el operador *lag* (Ver Lütkepohl, 2005, pp. 22).

En esta forma, Z_{n-1} la desviación del equilibrio al instante $n - 1$ es una variable explicativa, aunque, con un argumento recursivo, las desviación a distintos lags se pueden tomar en consideración, permitiendo así un ajuste gradual al equilibrio.

Bajo el supuesto de que las componentes del proceso X son $\mathrm{I}(0)$ o $\mathrm{I}(1)$, los procesos que aparecen en el modelo de corrección del error son todos estacionarios. Por lo tanto, es común realizar la predicción bajo este modelo y luego ajustar al nivel de la serie original (*integrando*), aunque este paso adicional incrementa la variabilidad al menos de manera lineal (similar a la varianza de una caminata aleatoria). Esto puede resultar en intervalos de predicción muy amplios y poco útiles.

Sin embargo, el concepto de cointegración en realidad proporciona una comprensión profunda de la relación entre las variables más que establecer un método directo de predicción.

El resultado que describe cómo un proceso X con componentes cointegradas puede escribirse como un modelo de corrección del error es conocido como el *Teorema de Representación de Granger*, que discutiremos a continuación.

Consideremos $\{X_n, n \in \mathbb{Z}\}$ como una serie de tiempo en $\mathbb{R}^m$ que sigue un modelo VAR. En adelante, X se

referirá al proceso $\{X_n, n \in \mathbb{Z}\}$, de manera similar se hará referencia a otros procesos estocásticos.

Si X es estable, entonces según el *Teorema de Descomposición de Wold*, X puede descomponerse como la suma de dos procesos no correlacionados $\{V_n, n \in \mathbb{Z}\}$ y $\{U_n, n \in \mathbb{Z}\}$:

$$X_n = V_n + U_n \quad \text{para todo } n \in \mathbb{Z},$$

donde V es un proceso determinístico que está completamente determinado por su pasado, y U es un proceso puramente estocástico, es decir, no determinístico. En adelante, nos referiremos a U como la parte estocástica del proceso X. Para una formalización detallada de estas definiciones se puede consultar la Sección 5.7 de Brockwell y Davis, 1991.

Supuesto 2.1

Supondremos que la serie de tiempo X estudiada no tiene componente determinístico. En la práctica, este componente puede eliminarse, por ejemplo, mediante la utilización de los residuos de una regresión de la forma:

$$X_n = a + bn + U_n,$$

Al realizar esta regresión, se asume que el componente determinístico es una función lineal. La validez de este supuesto se discute en la Sección 15.2 de Hamilton, 1994, aunque existen otras alternativas que pueden explorarse dependiendo del problema específico que se esté analizando.
Además, también suponemos que las series subyacentes no tienen un componente estacional. Para aprender cómo abordar este aspecto, se puede consultar la Sección 1.4 de Brockwell y Davis, 1991.

Finalmente, al Teorema de Descomposición de Granger es el siguiente:

Teorema 2.1

Sea $X \sim VAR(p)$ tal que $X \sim CI(1,1)$ y es un proceso puramente no determinístico. Supongamos que tiene la representación VECM

$$\Delta X_n := \alpha\beta^T X_{n-1} + \sum_{j=1}^{p-1} \Gamma_j \Delta X_{n-j} + \epsilon_n,$$

con $X_n = 0$ para $n < 0$, ϵ ruido blanco tal que $\epsilon_n = 0$ para $t \leq 0$.
Consideremos el polinomio

$$C(z) := (1-z)I_m - \alpha\beta^T z - \sum_{j=1}^{p-1} \Gamma_j (1-z) z^i.$$

Si se cumple que

1. *Si $\det C(z) = 0$ entonces $|z| \geq 1$.*
2. *El número de raíces unitarias es $m - r$.*
3. *$\alpha, \beta \in \mathbb{R}^{m \times r}$ tal que $rk\ \alpha = rk\ \beta = r$.*

Entonces

$$X_n = \mathfrak{F} \sum_{i=1}^{n} \epsilon_i + \mathfrak{F}^\star(L)\epsilon_n + X_0^\star,$$

donde

$$\mathfrak{F} = \beta_\perp \left[\alpha_\perp^T \left(I_m - \sum_{i=1}^{p-1} \Gamma_i \right) \beta_\perp \right]^{-1} \alpha_\perp^T,$$

entonces $rk\ \mathfrak{F} = m - r$, $\mathfrak{F}^\star(L)\epsilon_n \sim I(0)$ y $\beta_\perp, \alpha_\perp \in \mathbb{R}^{(m-r)\times r}$ son matrices cuyas columnas son ortogonales a las columnas de β y α, respectivamente.

Demostración. Ver Apéndice B. □

Observación 2.2

La variable $X_0^\star$ representa una constante distinta a la *condición inicial* X_0 de la serie X. Para más detalles sobre esta modificación, se puede consultar la prueba en el Apéndice B.

El Teorema de Representación de Granger es más general; no es estrictamente necesario que X siga un modelo VAR. Una versión más amplia se encuentra en Engel y Granger, 1987, y la primera prueba documentada correcta está en Johansen, 1991. La razón de introducir este teorema no es profundizar en el concepto de cointegración, sino utilizar estos resultados como punto de partida para motivar las metodologías desarrolladas en esta tesis.

2.2. Proyección a un espacio estable bajo la noción de cointegración

En principio, estamos ante una serie de tiempo X en $\mathbb{R}^m$ cuyas componentes son $\mathrm{I}(1)$ o $\mathrm{I}(0)$.

Este tipo de procesos son comunes en la práctica al analizar series temporales, especialmente en los campos de Economía y Finanzas, que constituyen la motivación principal detrás de estas ideas.

Supongamos que X sigue un modelo VAR y $X \sim \mathrm{CI}(1,1)$ con un rango de cointegración r, es decir, según las condiciones del Teorema de Representación de Granger. Observamos que este supuesto es restrictivo, dado que todas las componentes deben tener el mismo orden y además cointegrarse según la Definición 2.2.

Como corolario del Teorema de Representación de Granger, en el caso de un VAR inestable, se tiene una descomposición en $m-r$ *caminatas aleatorias*, que en términos económicos representan *tendencias comunes*, y r *componentes estacionarias*.

Una descomposición alternativa, consecuencia del Teorema de Representación de Granger, es la siguiente:

Proposición 2.1

Sea $\{X_n, n \in \mathbb{Z}\}$ una serie de tiempo en $\mathbb{R}^m$ que cumple las condiciones del Teorema 2.1. Se presenta una descomposición de la forma:

$$X_n = X_0^\star + P_{\beta_\perp} \cdot \left(\theta(1)\sum_{i=1}^{n}\epsilon_n\right) + P_\beta \cdot X_n + \eta_n \quad \text{para todo } n > 0, \tag{2.1}$$

donde $P_\beta, P_{\beta_\perp}$ representan las matrices de proyección a Col β y Col $\beta_\perp$, respectivamente, η es una serie de tiempo estacionaria, y

$$Z_n := P_\beta X_n + P_{\beta_\perp}\Delta X_n,$$

tiene una representación MA dada por

$$Z_n = \theta(B)\epsilon_n.$$

Demostración. Ver Apéndice B . □

Por construcción, las componentes de $P_\beta X_n$ son procesos $\mathrm{I}(0)$, lo que implica que la serie de tiempo $\{P_\beta X_n, n \geq 0\}$ es estacionaria.

La ventaja de esta proposición radica en que buscamos representar al proceso observado X en términos de una proyección hacia un espacio estable, conforme a las condiciones del Teorema 2.1, que corresponde al espacio generado por las relaciones de cointegración.

Por lo tanto, podemos desarrollar un esquema predictivo para una serie X en $\mathbb{R}^m$ con componentes $\mathrm{I}(1)$, centrándonos en predecir el proceso estacionario $\{P_\beta X_n + \eta_n, n > 0\}$.

Debido a que estamos realizando proyecciones, se minimiza el error, por lo que el esquema desarrollado para predecir X está basado en considerar predictores lineales en Col β, que bajo los supuestos del Teorema 2.1 se conoce como espacio de cointegración.

Como se mencionó anteriormente, el supuesto de que X cumple con las condiciones del Teorema 2.1 es restrictivo. Es plausible que las componentes de X simplemente no estén cointegradas. En este caso, si consideramos las m series, tendríamos $r = 0$, lo que significa que desde la perspectiva de la cointegración no existe un espacio estable al cual proyectar las m series en conjunto. Sin embargo, esto no implica que si consideramos dos subconjuntos separados de las componentes de X, cada uno visto como una serie de tiempo, puedan estar cointegrados individualmente. En consecuencia, podríamos realizar predicciones por separado para estos subconjuntos.

Lo anterior se debe a que el concepto de cointegración representa una relación entre todas las variables del sistema y una parametrización específica descrita por el modelo VECM.

Por lo tanto, lo ideal sería obtener relaciones de estabilidad a partir de una serie X con componentes inestables y no necesariamente cointegradas, que tengan capacidad predictiva y puedan manejar el hecho de que quizás no todas las variables deban ser consideradas para obtener la combinación lineal estable. Es decir, la carga asignada a esas variables en la combinación lineal puede ser 0. Esto, en principio, difiere del concepto de cointegración, ya que una relación de cointegración considera a priori que todas las variables tienen importancia para obtener la relación de estabilidad.

En lo sucesivo, describiremos algunas de las ideas detrás de las metodologías existentes para obtener relaciones de cointegración. Estas ideas servirán de motivación para estudiar alternativas más flexibles, como por ejemplo, la metodología propuesta en Muriel et al., 2012.

2.2.1. Procedimiento de Johansen

Uno de los métodos más utilizados en el análisis de cointegración es el denominado *procedimiento de Johansen*, introducido en Johansen, 1991. Una exposición resumida de este método se puede encontrar en la Sección 5.5.1 de Maddala y Kim, 1998.

Para aplicar este método, se supone que X cumple las condiciones del Teorema 2.1, sigue un modelo VAR de algún orden, y su proceso de innovación es gaussiano.

Supongamos que X tiene la representación VECM como en el Teorema 2.1 y sea

$$M_{p-1}(n) := \text{SPAN}\ \{\Delta X_{n-1}, \dots, \Delta X_{n-p+1}\}$$

Utilizando el Teorema de Proyección podemos obtener las descomposiciones

$$\Delta X_n = R_{0n} + P_{M_{p-1}(n)} \Delta X_n \in \mathbb{R}^m,$$

y

$$X_{n-1} = R_{1n} + P_{\mathcal{M}_{p-1}(n)} X_{n-1} \in \mathbb{R}^m,$$

donde $R_{0n} = P_{M_{p-1}(n)^\perp} \Delta X_n$ y $R_{1n} = P_{M_{p-1}(n)^\perp} X_{n-1}$. Equivalentemente, se pueden entender a R_{0n} y R_{1n} como los residuos de las regresiones entre ΔX_n y X_{n-1} contra las variables $\{\Delta X_{n-1}, \dots, \Delta X_{n-p+1}\}$.

Aplicando la proyección $P_{M_{p-1}(n)^\perp}$ a la parametrización VECM en el Teorema 2.1 obtenemos

$$R_{0n} = \alpha \beta^T R_{1n} + \epsilon_n.$$

Dada la muestra aleatoria

$$\mathbf{R}_0 = [\mathbf{r}_{01}^T, \dots, \mathbf{r}_{0T}^T]^T \in \mathbb{R}^{T \times m} \quad \text{y} \quad \mathbf{R}_1 = [\mathbf{r}_{11}^T, \dots, \mathbf{r}_{1T}^T]^T \in \mathbb{R}^{T \times m},$$

se puede estimar α y β resolviendo el problema de regresión multivariada múltiple suponiendo ϵ es un ruido gaussiano.

Sea

$$S := \begin{bmatrix} S_{00} & S_{10} \\ S_{10} & S_{11} \end{bmatrix},$$

donde

$$S_{ij} = T^{-1} \mathbf{R}_i^T \mathbf{R}_j \in \mathbb{R}^{m \times m} \quad \text{con } i, j = 0, 1.$$

Observación 2.3

En el Apéndice A de Johansen, 1991 se prueba que

- $\text{Var}(\beta^T R_{1n} \mid M_{p-1}(n)) \to \beta^T \Sigma_{11} \beta$ cuando $n \to \infty$.
- $\text{Var}(R_{0n} \mid M_{p-1}(n)) \to \Sigma_{00}$ cuando $n \to \infty$.
- $\text{Cov}(\beta^T R_{1n}, R_{0n} \mid M_{p-1}(n)) = \beta^T \Sigma_{10}$ cuando $n \to \infty$,

donde S_{ij} son las correspondientes versiones muestrales de Σ_{ij} para $i, j = 0, 1$.

Desde otro punto de vista, intentamos predecir linealmente un proceso $\text{I}(0)(\Delta X_n)$ con un proceso $\text{I}(1)(X_{n-1})$, por lo que intuitivamente buscamos las combinaciones lineales de X_{n-1} que estén más correlacionadas con ΔX_n.

En el procedimiento de Johansen se obtienen α y β por máxima verosimilitud suponiendo que ϵ es un ruido gaussiano.

Se observa que para β fijo

$$\hat{\alpha}(\beta) = (\beta^T S_{11} \beta)^{-1} \beta^T S_{10},$$

y la verosimilitud en función sólo de β es

$$|L(\beta)|^{-\frac{2}{T}} = \left| S_{00} - S_{01} \beta (\beta^T S_{11} \beta)^{-1} \beta^T S_{10} \right|,$$

en Johansen, 1991 se especifica que el estimador de máxima verosimilitud $\hat{\beta}$ es tal que

$$\hat{\beta} := \text{argmin}_{A \in \mathbb{R}^{m \times r}} \frac{\left| A^T (S_{11} - S_{10} S_{00}^{-1} S_{01}) X \right|}{|A^T S_{11} A|},$$

El valor máximo se obtiene a partir del problema de eigenvalores

$$S_{11}^{-1} S_{10} S_{00}^{-1} S_{01} v = \lambda v,$$

por lo que las columnas de $\hat{\beta}$ coincide con las primeras r correlaciones canónicas entre R_{1n} y R_{0n} (Ver Johnson, Wichern et al., 2002, Identidad 10-11) asintóticas pues no hay dependencia de n en estos eigenpares.

Observación 2.4

Observamos que este procedimiento supone de antemano:

1. Conocer el rango de cointegración r. Como se resume en Maddala y Kim, 1998, tener el rango de cointegración correcto es esencial para propósitos de predicción. En Johansen, 1991 se explica cómo estimar r utilizando un esquema de prueba de razón de verosimilitudes.
2. Asumir un modelo VAR, ya que se condiciona sobre el espacio $M_{p-1}(n)$. Además, como se argumenta en Harris, 1997, este procedimiento es sensible a la elección de p.

2.2.2. Metodología de Box-Tiao

Una manera alternativa de derivar las relaciones de cointegración utilizando correlación canónica es el *Método de Box-Tiao*.

Este método se basa en la idea del procedimiento paramétrico propuesto por Johansen (Johansen, 1991), pero intenta relacionar linealmente X_n con X_{n-1}.

Dado que ambas variables de respuesta son procesos con componentes $\text{I}(1)$ y estamos intentando predecir X_n, buscamos las combinaciones lineales de X_{n-1} menos correlacionadas con X_n, ya que, al final, se buscan componentes estacionarias.

La complementariedad existente entre estos dos enfoques de correlación canónica se debe a que los procesos $I(0)$ y $I(1)$ deben, en esencia, tener correlación cero, ya que el conjunto de procesos $I(d)$ es un espacio vectorial.

La forma de relacionar X_n y X_{n-1} es obtener residuales similares a R_{0n} y R_{1n}. El condicionamiento ya no sería respecto a $M_{p-1}(n)$. Estos residuales se obtienen haciendo la regresión de X_n y X_{n-1} contra un *componente determinístico* y una *dinámica a corto plazo*. La dinámica a corto plazo, en el procedimiento de Johansen, era el modelo autorregresivo de orden p.

Después de eliminar el componente determinístico y la dinámica a corto plazo, obtenemos, de forma similar al procedimiento de Johansen,

$$R_{0n} = \alpha\beta^T R_{1n} + \epsilon_n.$$

Lo que buscamos aquí son las correlaciones canónicas entre R_{0n} y R_{1n}. Con estos residuales, se define S de la misma forma.

Dado que ambos procesos son inicialmente $I(1)$, debemos considerar las r correlaciones canónicas más pequeñas. Así, las relaciones de cointegración son los r eigenvectores correspondientes.

Observación 2.5

Algo que hay que notar al trabajar con la matriz S es que esta se puede considerar como un estimador de la varianza a largo plazo del proceso X. La noción de cointegración implica una relación a largo plazo entre las variables.

2.2.3. Componentes Principales

La idea central de todas estas metodologías se concentra en analizar la dependencia temporal del proceso investigando la relación lineal entre X_n y X_{n-1}.

Motivados por estas ideas, en vez de utilizar correlación canónica, parece natural intentar descomponer la variabilidad explicada por combinaciones lineales de las componentes de X. Estos son los componentes principales.

De manera similar a la metodología de Box-Tiao, los componentes principales asociados a los r valores propios más pequeños corresponderán a estimadores de las relaciones de cointegración.

En Harris, 1997, se describe en detalle este método y los supuestos bajo los cuales los r eigenvectores asociados a los r eigenvalores más pequeños de la descomposición espectral de la matriz de covarianza muestral

$$S = T^{-1} \sum_{n=1}^{T} X_n X_n^T,$$

constituyen una base que estima, de forma consistente, el espacio vectorial generado por las relaciones de cointegración teóricas (Ver Harris, 1997, Lema 1).

Observamos que este método basado en PCA utiliza la descomposición de la varianza del proceso X en lugar de la dependencia entre las variables X_n y X_{n-1}. Así, PCA se basa en la idea de que las combinaciones lineales estables de un proceso con componentes no estacionarias deberían ser aquellas que acumulen la menor variabilidad.

La metodología propuesta en Harris, 1997, aunque hace ciertos supuestos, no requiere que el proceso X siga un modelo paramétrico. Por lo tanto, en principio, es más general y, para un proceso $X \sim CI(1,1)$, sería una metodología no paramétrica para la estimación de un conjunto de relaciones de cointegración.

Otra metodología similar que utiliza PCA, pero que en algún punto supone un modelo VAR, es el *método de Stock y Watson* (Ver Maddala & Kim, 1998, Sección 5.5.5). Este método se utiliza sobre todo en la tarea de identificar tendencias comunes, más que en la identificación de relaciones de cointegración.

Observación 2.6

Resumimos algunos de los puntos clave de todas estas metodologías, que están principalmente enfocadas en estimar una base para el espacio de cointegración:

1. El procedimiento de Johansen depende fuertemente del supuesto de ruido blanco gaussiano y es sensible al orden p. Es un método paramétrico.

2. El método por componentes principales no considera la dependencia entre las variables del proceso, la cual es natural en series de tiempo. PCA se basa en descomponer la varianza del proceso.

Finalmente, podemos concluir que, generalmente, bajo los supuestos del Teorema de Descomposición de Granger, analizar la dependencia lineal entre X_n y X_{n-1}, o ΔX_n y X_{n-1} en el caso del procedimiento de Johansen, permite estimar una base para el espacio de cointegración.

Esta conclusión es la motivación principal de esta sección. En lo sucesivo, explicaremos cómo se pueden utilizar estas ideas para obtener relaciones estables que, además, tengan capacidad predictiva, ya no necesariamente bajo las restricciones del Teorema de Representación de Granger.

2.3. PLS para obtener una proyección estable

2.3.1. Motivación

Como hemos enfatizado anteriormente, puede que estemos ante un proceso X que no cumpla los supuestos del Teorema 2.1, por ejemplo, una serie X con componentes de distinto orden de integración.

Motivado por el análisis de cointegración, se puede comenzar estudiando la dependencia entre las variables X_n y X_{n-1}. Así, se puede obtener un conjunto de componentes, concentradas en las columnas de una matriz β, tal que $\{P_\beta X_n, n \geq 0\}$ sea una serie de tiempo estacionaria. Esta serie se puede predecir y, con ello, entender la serie $\{X_n, n \geq 0\}$ subyacente.

Se puede utilizar la misma idea de que las componentes estables son precisamente aquellas que resultan de combinaciones lineales de las componentes de X con menor variabilidad.

Por otro lado, el propósito principal de este trabajo es abordar el problema de predicción. Por lo tanto, adicionalmente, se pueden buscar componentes que tengan capacidad predictiva. Esto es precisamente lo que hace el algoritmo PLS en el contexto de regresión.

En ese sentido, Muriel et al., 2012 propone una metodología basada en el algoritmo PLS para series de tiempo vectoriales. Esta metodología fue aplicada para estimar una base del espacio de cointegración. Sin embargo, como se menciona en Wold, 1979, el algoritmo PLS puede utilizarse en sistemas complejos con pocos supuestos teóricos.

Por lo tanto, se puede considerar que la metodología propuesta en Muriel et al., 2012 es independiente del análisis de cointegración, y su flexibilidad permite obtener componentes estables de sistemas con diversos tipos de estacionaridad, no exclusivamente sistemas con raíces unitarias.

Además, dado que la idea principal para obtener esas componentes estables es explorar la dependencia entre las variables X_n y X_{n-1}, aplicar el algoritmo PLS también nos proporcionaría componentes con capacidad predictiva.

El objetivo de esta sección es resumir las ideas presentadas en Muriel et al., 2012 y comparar esta metodología con las descritas anteriormente.

Observación 2.7

Algunas ventajas del algoritmo PLS son:

1. Tiende a utilizar el menor número de componentes necesario con capacidad predictiva, lo cual reduce la complejidad del modelo (Höskuldsson, 1988).

2. Considera la dependencia existente entre X_n y X_{n-1}, lo cual es natural en el contexto de series de tiempo.

3. Posee propiedades de ortogonalidad en las componentes (Garthwaite, 1994; Höskuldsson, 1988).

4. Es apropiado para modelos de alta dimensionalidad y grandes varianzas (Garthwaite, 1994), a diferencia del procedimiento de Johansen que puede enfrentar dificultades numéricas con la dimensionalidad, debido a la exhaustividad computacional del problema de máxima verosimilitud.

Estas ventajas hacen que el algoritmo PLS sea una alternativa a considerar en la exploración y análisis de relaciones de cointegración en series temporales.

La propuesta desarrollada en Muriel et al., 2012 se basa en la misma idea que las metodologías basadas en PCA y CCA. Además, PLS es más adecuado para el problema de predicción y representa una generalización de CCA. Por lo tanto, se puede considerar que PLS incorpora las ideas detrás de PCA y CCA para la estimación del espacio de cointegración, y en un contexto más amplio, para la estimación del espacio estable.

2.3.2. Descripción de la metodología

En lo siguiente, suponemos que $\{X_n, n \geq 0\}$ es una serie de tiempo en $\mathbb{R}^m$ con componentes no necesariamente estacionarias y que no necesariamente se cointegran. Es posible que haya componentes con integración de orden mayor (I(d) con $d > 1$), por lo tanto, X no necesariamente sigue un proceso cointegrado del tipo CI$(1,1)$.

La idea fundamental en Muriel et al., 2012 es determinar la matriz β de tal manera que

$$X_n = P_{\beta_\perp} X_n + P_\beta X_n \quad \text{para } n \geq 0,$$

donde el proceso $\{P_\beta X_n, n \geq 0\}$ es estacionario y $\{P_{\beta_\perp} X_n, n \geq 0\}$ no necesariamente lo es.

Dado que el proceso $\{P_\beta X_n, n \geq 0\}$ corresponde a una proyección del proceso original X, sabemos que entre las transformaciones lineales de X_n en Col β, la que minimiza el Error Cuadrático Medio (MSE) es precisamente la proyección. Por lo tanto, en principio, este proceso se puede utilizar como predictor de X.

La metodología en Muriel et al., 2012 se puede resumir en una forma de obtener β utilizando el algoritmo PLS.

Consideremos una muestra aleatoria $\{X_n, n = 1, 2, \ldots, T\}$ de la serie X en $\mathbb{R}^m$ con condición inicial X_0.

Sea

$$\mathbf{Y} := [X_T^T, \ldots, X_2^T]^T \in \mathbb{R}^{(T-1)\times m} \quad \text{y} \quad \mathbf{X} := [X_{T-1}^T, \ldots, X_1^T]^T \in \mathbb{R}^{(T-1)\times m}. \tag{2.2}$$

Aplicando el algoritmo PLS a la muestra aleatoria formada por **Y** como variable respuesta y **X** como variables explicativas se tiene para la primera iteración que

$$(\hat{w}_1, \hat{c}_1) := \underset{d\in\mathbb{R}^m, e\in\mathbb{R}^m}{\operatorname{argmax}} \operatorname{Cov}(\mathbf{X}d, \mathbf{Y}e)^2 \quad \text{con } \|d\|, \|e\| = 1, \tag{2.3}$$

según la notación del Algoritmo 1.2.

Naturalmente, si el proceso subyacente X tiene componentes no estacionarias, según lo discutido hasta ahora, esperaríamos que

$$T_1(n) := \hat{w}_1^T X_n \quad \text{para } n = 1, \ldots, T, \tag{2.4}$$

corresponda a una muestra aleatoria de una serie de tiempo no estacionaria. Esto se debe a que, según el Lema 2.1, el vector $\hat{w}_1$ corresponde al valor singular más grande de la matriz de covarianzas muestral, lo que implica que T_1 está altamente correlacionado con X.

Lema 2.1

Sea $\{X_n, n \in T\}$ una muestra aleatoria de la serie $X \in \mathbb{R}^m$. Sean **Y** *y* **X** *como en la ecuación (2.2). Considerando la primera iteración del algoritmo PLS en la ecuación (2.3), se cumple que*

$$\hat{c}_1 \propto \mathbf{Y}^T \mathbf{X} \hat{w}_1,$$

y por lo tanto, el problema de optimización en la ecuación (2.3) se puede simplificar a

$$\hat{w}_1 := \underset{d\in\mathbb{R}^m}{\operatorname{argmax}}\, d^T\mathbf{X}^T\mathbf{Y}\mathbf{Y}^T\mathbf{X}d \quad \textit{sujeto a } \|d\| = 1,$$

lo que implica que $\hat{w}_1$ es el eigenvector que corresponde al eigenvalor más grande de la matriz $\mathbf{X}^T\mathbf{Y}\mathbf{Y}^T\mathbf{X}$.

Demostración. Ver Apéndice B. □

Del Lema 2.1, se concluye que $\hat{w}_1$ es un vector correspondiente al valor singular más grande de la covarianza muestral, lo que implica que maximiza la capacidad predictiva.

Observación 2.8

Observamos que $T^{-1}\mathbf{Y}^T\mathbf{X}$ es un estimador muestral de

$$\Sigma_{10} := \lim_{n\to\infty} \operatorname{Cov}(X_n, X_{n-1}),$$

y de manera similar, $T^{-1}\mathbf{X}^T\mathbf{Y}$ es un estimador muestral de

$$\Sigma_{01} := \lim_{n\to\infty} \operatorname{Cov}(X_{n-1}, X_n).$$

Por lo tanto, según el Lema 2.1, tanto $\hat{w}_1$ como $\hat{c}_1$ pueden interpretarse como estimadores de w_1 y c_1, que corresponden a los problemas de eigenvalores:

$$\Sigma_{01}\Sigma_{10}w_1 = \lambda w_1,$$
$$\Sigma_{10}\Sigma_{01}c_1 = \lambda c_1.$$

Estamos suponiendo implícitamente que

$$T^{-1}\mathbf{Y}^T\mathbf{X} \xrightarrow{P} \Sigma_{10},$$
$$T^{-1}\mathbf{X}^T\mathbf{Y} \xrightarrow{P} \Sigma_{01},$$

sin embargo, esto solo ocurre si la relación de dependencia entre X_t y X_{t-1} es estable a largo plazo.

A pesar de la observación anterior, el estimador muestral de la covarianza es informativo a corto plazo, ya que indica cómo es, en promedio, la relación lineal entre variables consecutivas de la serie de tiempo.

Hasta el momento, hemos visto cómo el Lema 2.1 nos permite simplificar el problema de optimización que aparece en el Algoritmo 1.2. Se hace más énfasis en la componente $\hat{w}_1$ que en $\hat{c}_1$, ya que, en un contexto de predicción, es crucial obtener la dirección $\hat{w}_1$ asociada a la muestra aleatoria $\mathbf{X}$ para definir modelos de regresión similares a los presentados en la Sección 1.1.

En el caso de series de tiempo, tanto las variables contenidas en $\mathbf{Y}$ como en $\mathbf{X}$ toman valores en $\mathbb{R}^m$. Este es un caso particular que puede manejar el algoritmo PLS, donde se permite que las variables aleatorias subyacentes a $\mathbf{Y}$ y $\mathbf{X}$ tomen valores en espacios distintos.

Además, de la ecuación (2.2) se observa que $\mathbf{Y}$ y $\mathbf{X}$ se conforman del mismo proceso. Por lo tanto, en la siguiente iteración del algoritmo PLS, se espera obtener una muestra aleatoria de un proceso con base en el cual se puedan formar nuevas matrices $\mathbf{Y}$ y $\mathbf{X}$. Por lo tanto, en lugar de actualizar las matrices $\mathbf{Y}$ y $\mathbf{X}$ como en el Algoritmo 1.2, eliminamos la información contenida en la dirección $\hat{w}_1$ de cada observación de la muestra aleatoria $\{X_t, t = 1, \ldots, T\}$ de la siguiente forma:

$$\tilde{X}_n := \tilde{P}_{\hat{w}_1} X_n,$$

donde

$$\tilde{P}_{\hat{w}_1} := I_m - \frac{\hat{w}_1\hat{w}_1^T S}{\hat{w}_1^T S\hat{w}_1},$$

y

$$S = T^{-1} \sum_{n=1}^{T} X_n X_n^T,$$

es el estimador muestral de la matriz de covarianzas del proceso X.

Se observa que $\tilde{P}_{\hat{w}_1}$ es una matriz de proyección cuya imagen es $\text{SPAN}\{\hat{w}_1\}^{\perp}$. Aunque podríamos haber usado la matriz de proyección $I_m - \hat{w}_1 \hat{w}_1^T$, preferimos utilizar $\tilde{P}_{\hat{w}_1}$ en el Algoritmo 1.2 porque la matriz **X** se actualiza de la siguiente manera en la siguiente iteración:

$$\mathbf{X} \leftarrow \mathbf{X} \cdot \left(I_m - \frac{\hat{w}_1 \hat{w}_1^T \mathbf{X}^T \mathbf{X}}{\hat{w}_1^T \mathbf{X}^T \mathbf{X} \hat{w}_1} \right).$$

Además, $\tilde{P}_{\hat{w}_1}$ es una proyección ponderada por la varianza del proceso X.

Por lo tanto, la versión propuesta del algoritmo PLS discutida hasta ahora es la siguiente:

Algoritmo 2.1 (*Algoritmo PLS para series de tiempo multivariadas)*

Iniciamos con una muestra aleatoria centrada de una serie de tiempo $\{X_n^{(1)}, n = 1, \dots, T\}$ con valores en $\mathbb{R}^m$.
Sea **X** y **Y** como en (2.2) construidas a partir de $\{X_n^{(1)}, n = 1, \dots, T\}$
Para $i = 1, \dots, m$

1. Obtenemos $\hat{w}_i$ según el Lema 2.1.
2. Actualizamos la muestra aleatoria como

$$X_n^{(i+1)} \leftarrow \tilde{P}_{\hat{w}_i} X_n^{(i)}.$$

3. Se conforman nuevamente **X** y **Y** como en (2.2) con la muestra aleatoria en el numeral 2 y se hace $i \leftarrow i + 1$.

Notamos del Algoritmo 2.1 que iterativamente se reduce la varianza de la muestra $\{X_n, n = 1, \dots, T\}$ mediante componentes que también maximizan la dependencia lineal entre **Y** y **X**. Además, se espera que

$$T_i(n) := \hat{w}_i^T \tilde{X}_n^{(i)} \quad \text{para } n = 1, \dots, T,$$

corresponda a una serie de tiempo no estacionaria mientras que $\{\tilde{X}_n^{(i)}, n = 1, \dots, T\}$ sea un proceso no estacionario.

Sin embargo, puede haber un índice j para el cual por primera vez $\{X_n^{(j)}, n = 1, 2, \dots, T\}$ provenga de un proceso estacionario, momento a partir del cual las componentes $\hat{w}_j, \dots, \hat{w}_m$ darán lugar a muestras de series de tiempo univariadas $T_j, \dots, T_m$ estacionarias. Claro está, la existencia de dicho índice j depende del grado de no estacionaridad de la serie X.

En el contexto de cointegración, el siguiente resultado establece una conexión con la metodología descrita en Harris, 1997.

Lema 2.2

Sea X una serie de tiempo con valores en $\mathbb{R}^m$ que cumple las condiciones del Teorema de Representación de Granger y con proceso de innovación un ruido blanco fuerte. Sea $\{X_n, n = 1, \dots, T\}$ una muestra aleatoria de dicha serie.
Se cumple que el primer componente PLS $\hat{w}_1$ es asintóticamente equivalente, salvo signo, al primer componente principal de X, es decir, su diferencia converge en probabilidad a 0.
El primer componente principal corresponde al eigenvector asociado al eigenvalor más grande de la varianza muestral

$$S := T^{-1} \sum_{n=1}^{T} X_n X_n^T.$$

Demostración. Ver Apéndice B □

Observación 2.9

Un corolario del Lema 2.1 establece que $\hat{w}_1$ y $\hat{c}_1$ son asintóticamente equivalentes. Por lo tanto,

$$\tilde{P}_{\hat{w}_1} \approx I_m - \hat{w}_1 \hat{w}_1^T,$$

lo que sugiere que el método propuesto en Harris, 1997 y el basado en PLS deberían comportarse de manera similar cuando $X \sim \mathrm{CI}(1,1)$.

2.3.3. Identificación de las componentes estables

Anteriormente hemos observado que al aplicar el Algoritmo 2.1 a una serie $\{X_n, n = 1, \ldots, T\}$ en $\mathbb{R}^m$ obtenemos una colección de series

$$\{T_1, \ldots, T_m\},$$

donde T_i está definido en (2.4).

Si X es no estacionaria, según la naturaleza del Algoritmo 2.1, se espera que los procesos T_i sean no estacionarios hasta encontrar un índice j tal que el proceso $X^{(j)}$ sea estacionario. En tal caso, todas las series en el conjunto

$$\{T_j, \ldots, T_m\}$$

serán estacionarias. Esto permite formalizar la noción del espacio estable estimado.

Definición 2.4 (*Espacio estable estimado*)

Consideremos $\{X_n, n = 1, \ldots, T\}$ una muestra aleatoria de una serie en $\mathbb{R}^m$. Consideremos la notación del Algoritmo 2.1 y la definición de T_i en (2.4).
Sea

$$j^\star = \begin{cases} \text{mín}\{j \mid T_j \text{ es una serie estacionaria }\} & \text{si } \{T_j \mid T_j \text{ es una serie estacionaria }\} \neq \emptyset, \\ 0 & \text{de otra forma .} \end{cases}$$

Definimos el espacio estable estimado para la serie X como

$$\mathcal{H}^S := \begin{cases} \text{SPAN}\{\hat{w}_{j^\star}, \hat{w}_{j^\star+1}, \ldots, \hat{w}_m\} & \text{si } j^\star \neq 0 \\ \{0\} & \text{en otro caso ,} \end{cases}$$

y su dimensión la denotamos por

$$\hat{r} := \dim \mathcal{H}^S.$$

Observación 2.10

Algunas notas importantes sobre la definición anterior son las siguientes:

1. Si $\hat{r} = 0$, entonces el Algoritmo 2.1 no pudo encontrar componentes estables, lo que sugiere un alto grado de no estacionaridad.
2. Si el proceso X sigue una distribución $\mathrm{CI}(1,1)$, entonces $\hat{r}$ es un estimador no paramétrico del rango de cointegración.
3. Si $\hat{r} = m$, entonces hay evidencia de que el proceso subyacente X es estacionario.

Finalmente, la tarea siguiente para definir $\mathcal{H}^S$ es determinar el valor de $j^\star$ o, equivalentemente, el valor de $\hat{r}$, ya que

$$j^\star = \begin{cases} 0 & \text{si } \hat{r} = 0, \\ m - \hat{r} + 1 & \text{si } \hat{r} > 0. \end{cases}$$

Observamos que

$$H_0: \underbrace{\hat{r} > 0}_{T_m \text{ es I(0)}} \quad \text{vs} \quad H_1: \underbrace{\hat{r} = 0}_{T_m \text{ es I(1)}},$$

y la generalización de este esquema de prueba de hipótesis es

$$H_0: \underbrace{\hat{r} \geq r_0}_{T_{m-r_0} \text{ es I(0)}} \quad \text{vs} \quad H_1: \underbrace{\hat{r} < r_0}_{T_{m-r_0} \text{ es I(1)}}.$$

Se puede tomar la decisión de mantener la hipótesis de que $T_{p-r_0} \sim \mathrm{I}(0)$ o de rechazarla utilizando una prueba KPSS (Kwiatkowski et al., 1992), como se describe en Muriel et al., 2012.

En la implementación computacional, el valor de $j^\star$ se estima directamente. Utilizando la prueba KPSS, $j^\star$ será el primer índice j para el cual no se encuentra evidencia suficiente para rechazar la hipótesis de que $T_j \sim \mathrm{I}(0)$.

Esto no excluye la posibilidad de utilizar otras pruebas de hipótesis, como la prueba de Dickey-Fuller Aumentada, que es robusta para otros tipos de estacionaridad (Gonzalo & Lee, 2000). En este caso, el valor de $j^\star$ sería el primer valor para el cual se rechaza la hipótesis de que $T_j \sim \mathrm{I}(1)$.

Sería más apropiado emplear una prueba de hipótesis que evalúe cuándo T_j rechaza la hipótesis nula de no estacionaridad, en lugar de solo rechazar la existencia de una raíz unitaria. Esto implica una hipótesis nula más general. Por comodidad y debido al Supuesto 2.1, la razón de no estacionaridad considerada es el orden de integración.

La implementación computacional de este trabajo usa la prueba KPSS, esto para emular la metodología descrita en Muriel et al., 2012.

2.3.4. Esquema predictivo

Finalmente, describiremos el esquema predictivo desarrollado para la serie X después de haber estimado el espacio estable.

Sea $\hat{r}$ la dimensión del espacio estable. Si $\hat{r} = 0$, no hay nada que hacer: todas las componentes obtenidas por el Algoritmo 2.1 son no estacionarias y no pueden ser utilizadas. Por ejemplo, la metodología de predicción descrita en la Sección 1.2 no sería aplicable en este caso.

Si $\hat{r} > 0$, renombramos los vectores que conforman la base que genera el espacio estable, tenemos así

$$\mathcal{H}^S := \mathsf{SPAN}\{\hat{s}_1, \ldots, \hat{s}_{\hat{r}}\} \subset \mathbb{R}^m.$$

Se definen los scores estables

$$S_i(n) := \hat{s}_i^T X_n \quad \text{para } n = 1, \ldots, T \text{ y } i = 1, \ldots, \hat{r}.$$

Por construcción, S_i corresponde a una serie de tiempo estacionaria para cada $i = 1, \ldots, \hat{r}$.

Las componentes estables $\hat{s}_1, \ldots, \hat{s}_{\hat{r}}$ se construyen de manera que se maximiza la dependencia entre X_n y X_{n-1}.

Por lo tanto, se puede aproximar el proceso X utilizando las componentes $S_1, \ldots, S_{\hat{r}}$ mediante un modelo de regresión similar al descrito en la Sección 1.1, es decir,

$$X_i(n) = \alpha_{1i} S_1(n) + \alpha_{2i} S_2(n) + \cdots + \alpha_{\hat{r}i} S_{\hat{r}}(n) + \epsilon_n \quad \text{para } i = 1, \ldots, m.$$

con ϵ un ruido blanco fuerte.

Para la muestra aleatoria $\{X_n, n = 1, \ldots, T\}$, consideremos

$$\mathbf{X} = [X_1^T, \ldots, X_T^T]^T \in \mathbb{R}^{T \times m}.$$

Por otro lado, definimos

$$\beta := [\hat{s}_1, \dots, \hat{s}_{\hat{r}}] \in \mathbb{R}^{m \times \hat{r}},$$

y

$$\mathbf{S} := \mathbf{X}\beta \in \mathbb{R}^{T \times \hat{r}}. \tag{2.5}$$

El modelo de regresión propuesto en notación matricial es

$$\mathbf{X} = \mathbf{S}\alpha + \epsilon, \tag{2.6}$$

con $\alpha \in \mathbb{R}^{\hat{r} \times m}$ y $\epsilon \in \mathbb{R}^{T \times m}$ cuyas filas son la muestra aleatoria de un ruido blanco fuerte.

Por lo tanto, el modelo de regresión propuesto es una regresión multivariada clásica.

La estimación del modelo de regresión (2.6) permitirá reconstruir la información observada de la serie X a partir de las componentes **S**. De esta manera, se puede evaluar cuánta información se pierde de X al utilizar solo los componentes estables $S_1, \dots, S_{\hat{r}}$.

Si se utilizara el Algoritmo 2.1 para obtener los componentes $S_1, \dots, S_{\hat{r}}$, se esperaría que la pérdida de información fuera menor en comparación con otros métodos como el basado en PCA (Harris, 1997) o el procedimiento de Johansen.

De esta manera, los componentes estables $S_1, \dots, S_{\hat{r}}$ obtenidos con el algoritmo PLS deberían poder explicar mejor la dinámica entre las variables del sistema X.

Sea $\eta := [S_1, \dots, S_{\hat{r}}]$, la serie de componentes estables con valores en $\mathbb{R}^{\hat{r}}$, cuya muestra aleatoria obtenida a partir de $\{X_n, n = 1, \dots, T\}$ es **S**.

Por construcción, η es una serie de tiempo de menor dimensión compuesta por componentes estacionarias. Si suponemos que η sigue la dinámica de un modelo VAR, podemos establecer un esquema predictivo como el descrito en la Sección 1.2.

Las columnas de la matriz β representan los pesos de cada una de las variables de la serie X para formar las componentes de η.

Se puede estimar α del modelo (2.6) obteniendo su estimador de mínimos cuadrados (Ver Johnson, Wichern et al., 2002, Identidad 7-26).

Resumimos los pasos a seguir para el enfoque predictivo propuesto:

Algoritmo 2.2

Iniciamos con una muestra aleatoria $\{X_n, n = 1, \dots, T\}$ y un horizonte de predicción h.

1. Obtenemos el espacio estable estimado $\mathcal{H}^S = \text{SPAN}\{\hat{s}_1, \dots, \hat{s}_{\hat{r}}\}$. Se obtiene **S** como en (2.5).
2. Con base en la muestra aleatoria **S** se obtienen predictores puntuales $\hat{\eta}_{T+1}, \dots, \hat{\eta}_{T+h}$ con sus correspondientes intervalos de confianza, suponiendo un modelo VAR como se describe en la Sección 1.2.

Este algoritmo puede aplicarse de manera muy flexible para el ajuste y la predicción de series de tiempo vectoriales no estacionarias.

En la última sección de este capítulo se pondrá a prueba esta metodología en un análisis econométrico. Esto servirá para ilustrar muchas de las ideas descritas y demostrar que esta propuesta no se limita a series temporales vectoriales con variables cointegradas.

En el capítulo final de esta tesis, se explorará la generalización del Algoritmo 2.1 mediante el uso de elementos aleatorios funcionales. Se propondrá un esquema similar al Algoritmo 2.2 para definir un método predictivo en series de tiempo funcionales.

2.4. Análisis de variables asociadas a la Inflación en México de 2005 a 2023

Analizaremos el comportamiento de varias variables relacionadas con la inflación en México. Estas variables forman una serie de tiempo vectorial X, y están detalladas en la Tabla 1.

Variables	Descripción
P	Indice Nacional de Precios al Consumidor
U	Tasa de desocupación
BYM	Billetes y monedas
E	Tipo de cambio nominal
R	Tasa de interés interbancaria a 28 días
W	Salarios reales
PUSA	Indice Nacional de Precios al Consumidor de Estados Unidos

Tabla 1.: Variables que conforman la serie vectorial relacionado con la inflación en México

La información de cada una de las variables en la Tabla 1 abarca el periodo de febrero de 2005 a enero de 2023, con una frecuencia mensual. Esto proporciona una muestra aleatoria $\{X_n, n = 1, \dots, T\}$ donde $T = 216$ y $m = 7$.

Dado que la muestra aleatoria $\{X_n, n = 1, \dots, T\}$ cumple con el Supuesto 2.1, las visualizaciones que se presentarán a continuación mostrarán las series después de haber eliminado la tendencia lineal y los componentes estacionales correspondientes.

Por lo tanto, utilizaremos esta muestra aleatoria para estimar el espacio estable utilizando las siguientes metodologías descritas anteriormente:

1. Johansen.
2. PLS, según el Algoritmo 2.1.
3. PCA.

Se utiliza la implementación del procedimiento de Johansen mediante la función ca.jo de la librería urca en R (Pfaff, 2008). Es importante destacar que el procedimiento de Johansen estima los parámetros por máxima verosimilitud, lo que limita computacionalmente el valor de m; de hecho, la función ca.jo solo puede estimar hasta $m < 12$.

En este caso particular, dado que $m = 7$, podemos implementar las tres metodologías para estimar el espacio estable. Posteriormente, se realizarán los pronósticos según el Algoritmo 2.2.

El horizonte de predicción para el cual se realizarán los pronósticos es $h = 5$, lo que significa que se generarán pronósticos para los componentes $S_1, \dots, S_{\hat{r}}$ cinco meses hacia adelante.

Para evaluar qué tan bien se realiza la predicción con $h = 5$, se estima el espacio estable utilizando la muestra aleatoria $\{X_n, n = 1, \dots, T - h\}$. Posteriormente, se obtienen pronósticos de $S_1, \dots, S_{\hat{r}}$ para $n = T - h + 1, \dots, T$. Con estos pronósticos, se pueden obtener las estimaciones $\hat{X}_{T-h+1}, \dots, \hat{X}_T$ mediante la relación (2.6) y se comparan con los valores observados $X_{T-h+1}, \dots, X_T$.

En primer lugar, analizamos la estacionaridad de cada una de las series en la Tabla 1. Aplicando la prueba KPSS, cuya hipótesis nula es la estacionaridad de la serie correspondiente, se obtuvieron los siguientes p-valores para cada serie:

Variable	p-valor
P	0.01
U	0.02
BYM	0.10
E	0.09
R	0.02
W	0.01
PUSA	0.06

Tabla 2.: p-valores de la prueba KPSS aplicada a las muestras aleatorias de cada una de las variables en la Tabla 1

Por lo tanto, al observar la Tabla 2, se evidencia que la muestra aleatoria X proviene de una serie de tiempo con componentes no estacionarias. Con la excepción de las series correspondientes a las variables BYM (en escala logarítmica) y PUSA, utilizando un nivel de significancia de $\alpha = 0.05$, las demás componentes son no estacionarias y podrían tener un orden de integración $d \geq 1$, el cual no necesariamente es el mismo para todas ellas.

Precisamente, bajo ese escenario el procedimiento de Johansen no es apropiado porque no se tiene un proceso CI$(1,1)$, sin embargo, lo aplicamos para ver a que resultados conduce.

Método	Valor de $\hat{r}$
Johansen	3
PLS	5
PCA	5

Tabla 3.: Estimadores de la dimensión del espacio estable estimado utilizando los distintos métodos propuestos y la muestra aleatoria $\{X_1, \ldots, X_{T-h}\}$ con $h = 5$ para la estimación

En un sistema cointegrado, esperaríamos que los estimadores de $\hat{r}$ sean iguales o al menos valores cercanos. Los resultados de la Tabla 3 indican que la serie de tiempo subyacente a la muestra aleatoria $\{X_t, t = 1, \ldots, T\}$ no es cointegrada. Además, esto sugiere que el procedimiento de Johansen podría estar arrojando resultados espurios en este contexto.

Para evaluar cuánta información se pierde, se estima el modelo de regresión (2.6) para cada método. Luego, se visualiza la aproximación de X mediante las componentes estables y se reporta el error cuadrático medio de la aproximación en la Tabla 4. El error de estimación se refiere al error cuadrático medio en la aproximación respecto a la muestra $\{X_1, \ldots, X_{T-h}\}$. El error de predicción se calcula comparando con las observaciones $\{X_{T-h+1}, \ldots, X_T\}$.

Notamos que el procedimiento de Johansen no captura la dinámica de la serie, probablemente debido a que los supuestos necesarios para su aplicación no se cumplen. Además, la dimensión del espacio estable obtenida es menor en comparación con los otros dos métodos, lo que naturalmente reduce su poder predictivo.

En contraste, para las aproximaciones utilizando PCA y PLS, los ajustes se pueden observar en la Figura 2 y la Figura 3, respectivamente. Los errores cuadráticos medios (MSE) de las aproximaciones se resumen en

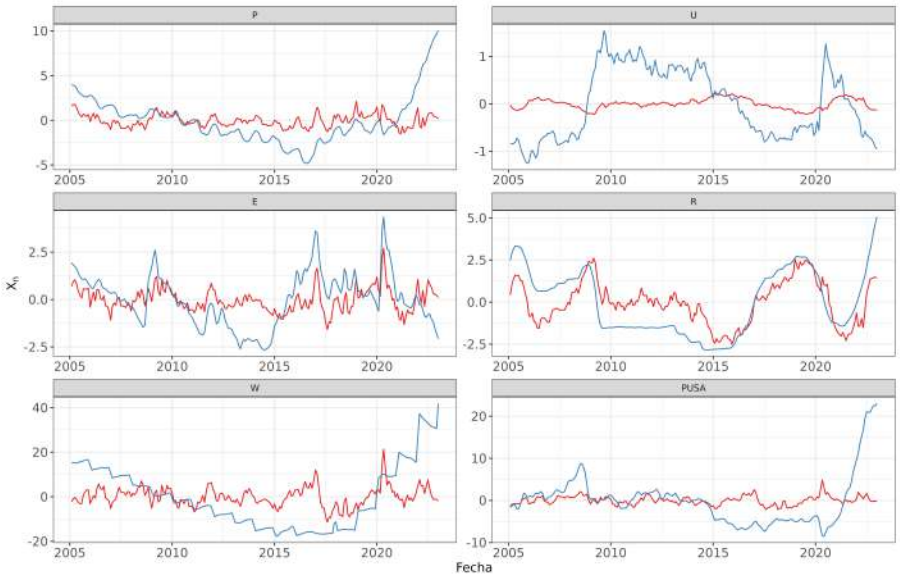


Figura 1.: Aproximación de las series en la Tabla 1 utilizando el procedimiento de Johansen para obtener el espacio estable. En línea sólida azul, se muestra el valor observado de cada serie. En línea roja, la aproximación usando la regresión en (2.6).

la Tabla 4.

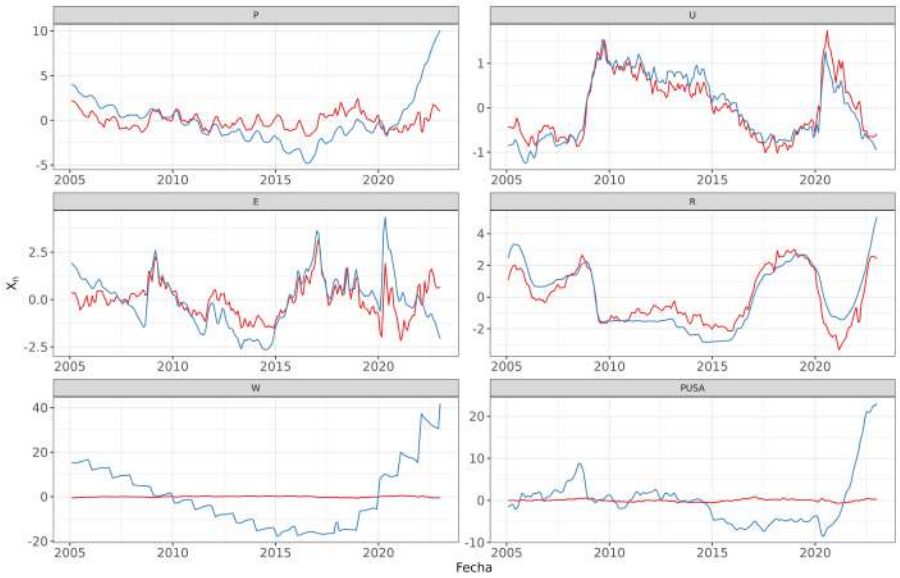


Figura 2.: Aproximación de las series en la Tabla 1 utilizando PCA para obtener el espacio estable. En línea sólida azul se muestra el valor observado de cada serie mientras que en línea roja la estimación-predicción con el método correspondiente.

Podemos observar que el algoritmo PLS para obtener el espacio estable pierde menos información, tanto en la estimación como en la predicción, en comparación con el procedimiento de Johansen y PCA.

Es interesante notar que PCA es igual de flexible que la propuesta basada en PLS, pero no exhibe el mismo poder predictivo. Esto sugiere que la función objetivo utilizada para obtener la base del espacio estable es fundamental. Aunque los espacios estables $\mathcal{H}^S$ obtenidos por PLS y PCA pueden ser los mismos, la naturaleza de la base estimada puede diferir significativamente.

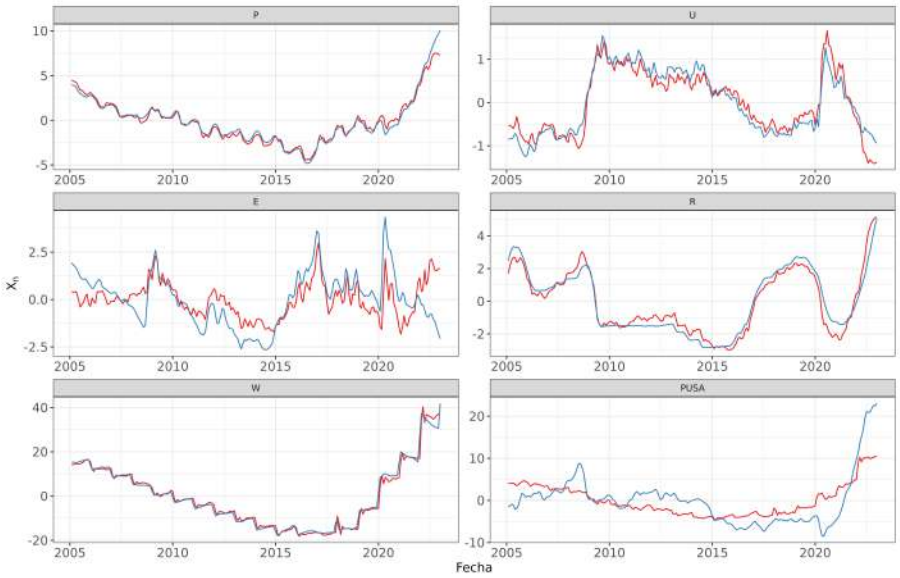
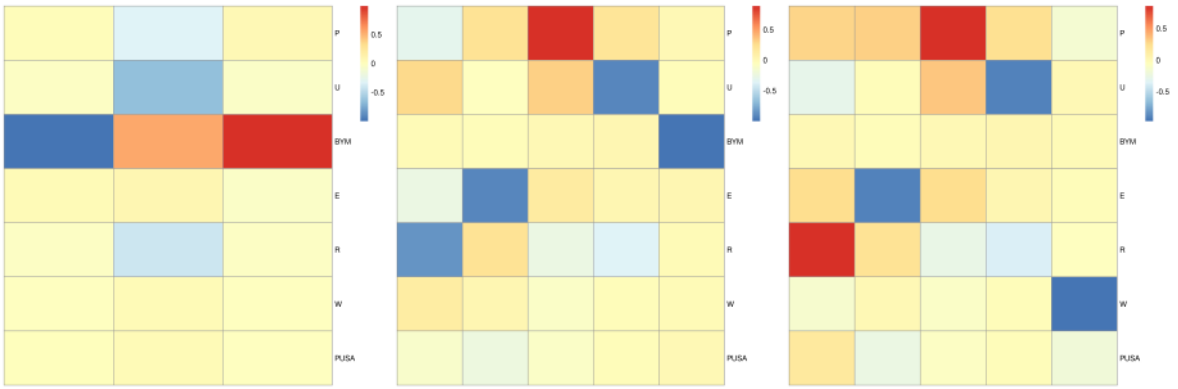


Figura 3.: Aproximación de las series en la Tabla 1 utilizando PLS para obtener el espacio estable. En línea sólida azul se muestra el valor observado de cada serie mientras que línea roja la aproximación obtenida con la regresión (2.6).

Método	MSE (estimación)	MSE (predicción)
Johansen	11.72	41.98
PLS	3.47	13.45
PCA	12.38	41.08

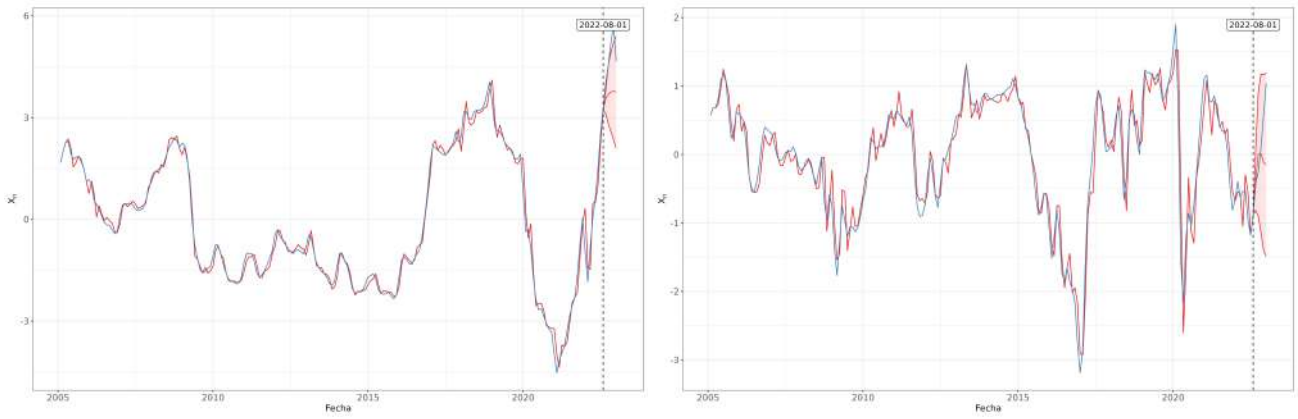
Tabla 4.: Error cuadrático medio de estimación y de predicción en la aproximación. Se usa un horizonte de predicción $h = 5$ meses para cada uno de los métodos.



(a) Johansen (b) PCA (c) PLS

Figura 4.: Magnitudes de las cargas $\hat{s}_1, \ldots, \hat{s}_{\hat{r}}$ para cada método. Cada entrada del vector $\hat{s}_i$ representa el peso que tiene la variable en la componente estable

Visualizamos las predicciones de las componentes $S_1, \ldots, S_{\hat{r}}$ obtenidas mediante el algoritmo PLS. En la Figura 5 se presenta el resultado específico para los primeros dos scores, S_1 y S_2.



(a) Primer score (b) Segundo score

Figura 5.: Predicción de las primeras dos scores estables S_1 y S_2 obtenidos con PLS. En azul, el valor observado y en rojo la predicción. El horizonte de predicción es $h = 5$ meses y el intervalo de predicción es del 80 %

Asimismo, en la Figura 5 se presentan las predicciones puntuales y por intervalo de los scores. Se observa que la amplitud del intervalo de predicción no crece exponencialmente debido a la estacionaridad. Además, se indica la fecha a partir de la cual se realizaron los pronósticos.

Finalmente, en la Tabla 1 se muestran las cargas de cada una de las variables que conforman las componentes $S_1, \ldots, S_{\hat{r}}$ para cada método utilizado.

En el procedimiento de Johansen, las magnitudes de las cargas se pueden visualizar en la Figura 4a. Se observa que las cargas están principalmente concentradas en la serie BYM, la cual se había identificado como estacionaria. Este resultado sugiere que el procedimiento de Johansen puede generar componentes estables de forma trivial cuando hay una serie estacionaria dentro del sistema. Sin embargo, estas componentes no resultan informativas, ya que no están relacionadas con las otras variables del sistema.

En contraste, PLS y PCA proporcionan cargas más informativas en comparación con la opción trivial elegida por Johansen. Las cargas estimadas por PLS y PCA son bastante similares, excepto por la primera componente, que es la más significativa debido a su capacidad predictiva superior por diseño. La diferencia entre la primera componente obtenida por PLS y PCA radica únicamente en el signo, lo cual condiciona la interpretación. Dado que la base obtenida por PLS aproxima mejor la muestra observada, consideramos que el signo obtenido mediante PLS es el correcto.

En conclusión, el procedimiento de Johansen tiende a fallar si no se aplica cuando $X \sim \text{CI}(d, b)$, típicamente con $d = 1$ y $b = 1$. Tanto PLS como PCA producen resultados bastante similares, y se podría investigar una relación más formal como la del Lema 2.2 bajo hipótesis más generales. El signo de las cargas es crucial para la interpretación precisa. Dado que PLS recupera mejor la información original, las cargas obtenidas mediante PLS, incluyendo sus signos, deberían considerarse más confiables.

§ 3. Componentes estables de una serie de tiempo funcional

En este capítulo, investigaremos ideas análogas a las discutidas en capítulos previos, pero en el contexto de series de tiempo funcionales.

En la primera parte, abordaremos el concepto de *cointegración funcional*, que, al igual que en el Capítulo 2, servirá como punto de partida para definir un esquema predictivo basado en componentes estables, aplicado a una serie de tiempo funcional no estacionaria.

En la segunda parte, propondremos una generalización del Algoritmo 2.1 para series de tiempo funcionales, constituyendo uno de los resultados principales de esta tesis.

Finalmente, implementaremos un esquema predictivo similar al propuesto en el Algoritmo 2.2 para el caso funcional y lo evaluaremos mediante un estudio de simulaciones.

3.1. Introducción a la concepto de cointegración funcional

El concepto de cointegración discutido en el Capítulo 2 sirvió como punto de partida para analizar los componentes estables de una serie de tiempo vectorial. En ese contexto, la serie de tiempo vectorial se considera un proceso estocástico sobre un espacio vectorial de dimensión finita.

Ahora, abordaremos el concepto de cointegración en un contexto más general: una serie de tiempo funcional, es decir, una serie que toma valores en un espacio de Hilbert.

Naturalmente, se deben hacer ciertos supuestos para poder tratar con procesos en espacios de funciones. En el caso de dimensión finita, los supuestos necesarios para hablar de cointegración involucran la definición misma y las propiedades descritas en el Teorema 2.1.

Para el caso funcional o de dimensión infinita, se espera que estos supuestos también estén comprendidos en lo que sería una generalización del Teorema 2.1 para espacios de Hilbert.

Antes de discutir dicha generalización, describiremos algunos de los supuestos necesarios para llegar a ese resultado, los cuales, al igual que en el caso multivariado, definirán el contexto de cointegración.

En primer lugar, enunciaremos la noción de un proceso $\mathrm{I}(d)$ para el caso funcional. Esta noción varía de autor a autor, la presentada en Beare et al., 2017 es la siguiente:

Definición 3.1

Decimos que una serie de tiempo funcional $\{X_n, n \in \mathbb{Z}\}$ es $\mathrm{I}(0)$, que lo denotamos por $X \sim \mathrm{I}(0)$, si es estacionaria y con un operador de covarianza a largo término no cero.

Decimos que X tiene orden de integración d, que lo denotamos por $X \sim \mathrm{I}(d)$, si X es $\mathrm{I}(0)$ luego de diferenciar d veces.

En el Teorema 2.1 se discute el caso $\mathrm{CI}(1,1)$, que es muy común en la práctica. Para realizar un análisis similar, consideraremos en lo que resta de la sección que $X \sim \mathrm{I}(1)$. Aunque, al igual que en dimensión finita, el Teorema de Representación de Granger es más general, también existen versiones más amplias para el caso funcional. Por ejemplo, un teorema de representación para series de tiempo funcionales $\mathrm{I}(2)$ (Beare & Seo, 2020), pero esto sigue siendo un problema de investigación actual.

Supongamos entonces $X \sim \mathrm{I}(1)$ bajo la Definición 3.1. Además, supongamos que ΔX tiene la representación MA

$$\Delta X_n = \sum_{k=0}^{\infty} A_k(\epsilon_{n-k}) \quad \text{para } n \geq 0, \tag{3.1}$$

con $\{\epsilon_n, n \in \mathbb{Z}\}$ un ruido blanco fuerte en $L^2_{\mathcal{H}}$.

De ahora en adelante, consideramos una serie $\{X_n, n \geq 0\}$, es decir, con condición inicial.

El primer supuesto es que $\sum_{k=0}^{\infty} \|A_k\| < \infty$. Según lo especificado en la Observación 1.15, esto implica que el operador $A := \sum_{k=0}^{\infty} A_k$ está bien definido en $\mathcal{L}$. El operador de covarianza a largo término del proceso ΔX viene dado por

$$\Gamma_{\Delta X} := AC_\epsilon A^\star,$$

donde C_ϵ es el operador de covarianza de cualquiera de las variables del proceso ϵ, que suponemos positivo definido. Además, se resalta con el subíndice el proceso considerado para obtener el operador de covarianza a largo término.

Hasta este momento, solo hemos formalizado la idea de que el proceso X debe ser $\mathrm{I}(1)$ y que ΔX tenga una representación MA bien definida.

El supuesto adicional clave a través del cual Beare et al., 2017 ofrece una definición del concepto de cointegración para este caso, y generaliza las ideas del Teorema 2.1, es el siguiente.

Supuesto 3.1

Sea $\{X_n, n \geq 0\}$ una serie de tiempo funcional con $X \sim \mathrm{I}(1)$, tal que ΔX tiene una representación MA bien definida como en (3.1), con un proceso de innovación ϵ y un operador de covarianza C_ϵ positivo definido. Supongamos además que

$$\sum_{k=0}^{\infty} k\|A_k\| < \infty.$$

Finalmente, definiremos el espacio de cointegración y explicaremos por qué puede considerarse una componente de estabilidad.

Definición 3.2

Sea $\{X_n, n \in \mathbb{Z}\}$ una serie de tiempo funcional con $X \sim \mathrm{I}(1)$ según la Definición 3.1. Llamamos espacio de cointegración de X al subespacio $\ker \Gamma_{\Delta X}$. Denotamos este espacio, sobreentendiendo el proceso subyacente, como $\mathcal{H}^S := \ker \Gamma_{\Delta X}$, y definimos el espacio atractor como su complemento ortogonal, es decir, $\mathcal{H}^N := (\ker \Gamma_{\Delta X})^\perp$.

Precisamente, bajo las condiciones de su existencia, los elementos del espacio de cointegración son componentes de estabilidad, como se indica en la siguiente proposición.

Proposición 3.1

Bajo el Supuesto 3.1, $x \in \mathcal{H}$ pertenece a $\mathcal{H}^S$ si y solo si la serie $\{\langle X_n, x\rangle, n \geq 0\}$ es estacionaria.

Demostración. Ver Proposición 3.1 de Beare et al., 2017 □

Observación 3.1

Aunque la Proposición 3.1 de Beare et al., 2017 toma en cuenta la condición inicial X_0, es importante recordar que el concepto de estacionaridad, tanto en el caso de dimensión finita como en el funcional, es asintótico. Esto significa que a medida que transcurre el tiempo, la influencia de la condición inicial tiende a anularse.
Por lo tanto, en adelante consideraremos, sin pérdida de generalidad, $X_0 = 0$, lo cual simplifica el enunciado de la proposición anterior.

Por otro lado, consideramos una relación de estabilidad en el siguiente sentido.

Definición 3.3

Decimos que $x \in \mathcal{H}$ es una relación de estabilidad para la serie funcional X si la serie univariada $\{\langle X_n, x\rangle, n \geq 0\}$ es estacionaria. Denotamos al espacio estable como $\mathcal{H}^S$ y a su complemento ortogonal como $\mathcal{H}^N$, que llamaremos espacio de inestabilidad.

Por la Proposición 3.1, sabemos que bajo el Supuesto 3.1, el espacio estable coincide con el espacio de cointegración, por lo tanto, no hay ambigüedad.

Sin embargo, la noción de relación o dirección de estabilidad es más general, ya que no está restringida al Supuesto 3.1, aunque en este caso existe una forma analítica para $\mathcal{H}^S$ como kernel de un operador.

Por lo tanto, al igual que se discutió en el Capítulo 2, se pretende emular los Algoritmos 2.1 y 2.2 para el caso funcional, con la esperanza de que sean igual de flexibles y proporcionen un esquema útil incluso fuera de los supuestos inherentes al concepto de cointegración funcional.

Finalmente, mientras que la generalización del Teorema 2.1 se encuentra en el Teorema 4.1 de Beare et al., 2017, para el desarrollo de este trabajo resulta más útil la *Descomposición de Beveridge-Nelson*, que puede considerarse como el análogo funcional de la Proposición 2.1.

Proposición 3.2 (*Beveridge-Nelson*)

Sea $\{X_n, n \geq 0\}$ una serie de tiempo funcional que cumple con el Supuesto 3.1. Definamos $\tilde{A}_k = -\sum_{j=k+1}^{\infty} A_j$ para $k \geq 0$. Entonces la serie

$$\eta_n := \sum_{k=0}^{\infty} \tilde{A}_k \epsilon_{n-k},$$

es convergente en $L^2_{\mathcal{H}}$ y es una serie estacionaria. Además, se tiene que

$$X_n = A\left(\sum_{i=0}^{n} \epsilon_i\right) + \eta_n.$$

De manera análoga, el esquema predictivo se basará en realizar la predicción sobre el proceso estacionario η, que es la proyección de X en el espacio estable $\mathcal{H}^S$.

Por otro lado, uno de los modelos utilizados en series de tiempo funcionales es el modelo autorregresivo funcional (Ver Bosq, 2000, Capítulo 3), en el cual se pueden ilustrar los conceptos presentados hasta el momento.

Ejemplo 3.1 (*AR Funcional*)

Consideremos el modelo autorregresivo funcional dado por

$$X_n = \rho(X_{n-1}) + \epsilon_n,$$

donde ϵ es un ruido blanco fuerte en $\mathcal{L}^2_{\mathcal{H}}$.
Si ρ es un operador compacto, entonces se satisface el Supuesto 3.1.

Al igual que el estudio del concepto de cointegración en el caso multivariado sirvió para introducir las ideas de una proyección a un espacio estable, también ha resaltado que la noción de este espacio estable es más general. En las secciones siguientes, se describirán las metodologías para estimar este espacio y se formalizará cómo puede ser utilizado para la predicción de una serie de tiempo funcional estacionaria.

3.2. Proyección a un espacio estable para una serie de tiempo funcional

En el caso de dimensión finita, existen varios procedimientos para la estimación del espacio de cointegración, como el procedimiento de Johansen, Box-Tiao y PCA. En contraste, en el caso funcional, generar metodologías análogas es un tema de investigación actual. Por ejemplo, en Seo, 2023a, se generaliza el procedimiento de Harris, 1997 bajo la idea de cointegración funcional discutida en la sección anterior.

El objetivo principal es estimar el espacio estable $\mathcal{H}^S$, que, según el Supuesto 3.1, equivale a estimar una base del espacio de cointegración.

Una vez estimada la base, es directo construir la proyección en los espacios correspondientes.

Observación 3.2

Dado que $\mathcal{H}^N \subset \mathcal{H}$ es cerrado, la proyección $P^N \in \mathcal{L}$ a este espacio existe y es única. En consecuencia, la proyección P^S a $\mathcal{H}^S$ también existe y es única, y según el Teorema de Proyección se define como

$$P^S := I - P^N,$$

donde I es el operador identidad en $\mathcal{L}$.

Para que la proyección en $\mathcal{H}^S$ sea informativa, es necesario que el grado de estacionaridad no sea tan severo. Este concepto se formaliza en el siguiente supuesto.

Supuesto 3.2

Sea X una serie de tiempo funcional con $X \sim \mathrm{I}(1)$. Suponemos que

$$\varphi := \dim \mathcal{H}^N < \infty.$$

En el contexto de cointegración, esto equivale a

$$\dim((\ker A)^{\perp}) = \dim \operatorname{cl} \operatorname{ran} A^{\star} \leq \infty,$$

lo que significa que $A^{\star}$ tiene rango finito. La identidad

$$(\ker A)^{\perp} = \operatorname{cl} \operatorname{ran} A^{\star},$$

se conoce como el *Teorema de la Dimensión Fuerte* (Ver Beare et al., 2017; Conway, 2019).

En Seo, 2023a, se menciona que este supuesto es común en la literatura reciente sobre el análisis de series de tiempo funcionales.

El objetivo será encontrar una base de $\mathcal{H}^N$ para obtener un estimador $\hat{P}^N$ de P^N. En consecuencia, el estimador de P^S sería

$$\hat{P}^S := I - \hat{P}^N.$$

Al igual que en el Capítulo 2, se discutirán distintas metodologías existentes en la literatura para la estimación de estas proyecciones.

En el contexto funcional, no existe un procedimiento análogo al de Johansen, pero sí existe un método basado en PCA (Seo, 2023a), el cual puede entenderse como una generalización del método propuesto en Harris, 1997 para el caso funcional.

En lo sucesivo, se discutirá la metodología presentada en Seo, 2023a y se describirá la propuesta presentada en esta tesis, la cual generaliza el Algoritmo 2.1 para el caso funcional.

3.2.1. Componentes Principales Funcionales

La metodología propuesta en Seo, 2023a para estimar las proyecciones a los espacios estables se basa en FPCA (Análisis de Componentes Principales Funcionales).

Sea $\{X_n, n = 1, \ldots, T\}$ una muestra de una serie de tiempo funcional con media 0 y, sin pérdida de generalidad, consideremos $X_0 = 0$.

Consideremos el operador de covarianza muestral $\hat{C}$ con kernel $\hat{c}$ dado por

$$\hat{c}(t,s) = T^{-1} \sum_{i=1}^{T} X_i(t) X_i(s).$$

Sean $\{\hat{\lambda}_j, j \in \mathbb{N}\}$ y $\{\hat{v}_j, j \in \mathbb{N}\}$ el conjunto de eigenvalores y eigenvectores del operador $\hat{C}$, respectivamente, que corresponden a los scores y componentes principales estimados para el proceso X.

Bajo el Supuesto 3.2 y suponiendo que φ es conocido, los estimadores $\hat{P}^N$ y $\hat{P}^S$ son dados por

$$\hat{P}^N_{\mathrm{FPCA}} := \sum_{j=1}^{\varphi} \hat{v}_j \otimes \hat{v}_j \quad \text{y} \quad \hat{P}^S_{\mathrm{FPCA}} := I - \hat{P}^N_{\mathrm{FPCA}}. \tag{3.2}$$

En Chang et al., 2016, se establece que $\|\hat{P}^N - P^N\|_{\mathcal{L}} = O_p(T^{-1})$ bajo el Supuesto 3.1 y otras restricciones (Ver Seo, 2023a, Supuesto W).

La Sección 3.2 de Seo, 2023a se discute un modificación de la estimación (3.2) obteniendo estimadores más eficientes, pues se remueven los parámetros desconocidos en la distribución asintótica de $T(\hat{P}^N - P^N)$ (Ver Seo, 2023a, Teorema 3.1). Este modificación es análogo al presentado en Harris, 1997.

Sin embargo, dicha modificación está sujeta al Supuesto 3.1 y enfocada a mejorar la estimación de φ.

La modificación no está necesariamente condicionada al uso de FPCA y se basa en la manera en que se estima el operador de covarianza a largo plazo según Horváth y Kokoszka, 2012. Se podría explorar esta modificación en la generalización del Algoritmo 2.1 propuesto. Por el momento, dejaremos de lado esta modificación y compararemos los estimadores en (3.2) con los propuestos en la siguiente sección.

El supuesto de conocer φ no es realista. Por lo tanto, también describiremos métodos para estimarlo, ya sea utilizando FPCA o el método que presentaremos.

3.3. Mínimos Cuadrados Parciales Funcionales

Como se discutió en el Capítulo 1, en el caso multivariado, varios estudios han desarrollado las propiedades teóricas del Algoritmo PLS aplicado a modelos de regresión, donde tanto la variable respuesta como la explicativa se encuentran en un espacio de dimensión finita.

En el contexto funcional, también se han explorado trabajos que intentan generalizar el Algoritmo PLS cuando algunas de las variables, ya sea la respuesta o la explicativa, toman valores en un espacio de Hilbert. Dos perspectivas destacadas en esta línea de investigación son:

1. La perspectiva presentada en Preda y Schiltz, 2011 reduce el Algoritmo PLS a un problema de eigenvalores utilizando los *operadores de Escoufier* (Escoufier, 1970). Se trata de un enfoque para problemas de regresión donde tanto la variable respuesta como la explicativa son funcionales, lo cual representa una alternativa plausible. Este enfoque se aplica, por ejemplo, en Preda et al., 2007 para resolver el problema de discriminante lineal.

2. Un enfoque detalladamente descrito se encuentra en Delaigle y Hall, 2012. Aquí se presentan dos alternativas: una que construye componentes ortogonales y otra con mejores propiedades asintóticas. Este enfoque se aplica especialmente cuando la variable respuesta es escalar.

La alternativa en Delaigle y Hall, 2012 es mucho más completa; sin embargo, no se puede aplicar directamente para generalizar el Algoritmo 2.1, ya que en nuestro caso tanto las variables respuesta y las explicativas son funcionales.

Por otro lado, el problema de eigenvalores descrito en Preda y Schiltz, 2011, a través de los operadores de Escoufier, no parece ser una generalización directa del problema de eigenvalores en el Lema 2.1.

Por lo tanto, propondremos una generalización del Lema 2.1 y, por ende, del Algoritmo 2.1, basada en los operadores de covarianza cruzada.

Para realizar la generalización del algoritmo PLS, utilizaremos como base los resultados presentados en la Sección 1.1, enfocándonos en particular en la Proposición 1.2 y el Lema 2.1.

Sean Y y X dos elementos aleatorios en $L^2_{\mathcal{H}_Y}$ y $L^2_{\mathcal{H}_X}$, respectivamente, ambos con media cero. Los espacios de Hilbert $\mathcal{H}_Y$ y $\mathcal{H}_X$, donde toman valores Y y X, respectivamente, no necesariamente son el mismo.

Así como en la Sección 1.1, se entiende a Y como la variable respuesta que se busca explicar mediante la variable explicativa X.

El problema de optimización en la Proposición 1.2 en el caso funcional es:

$$(w_1, c_1) = \underset{d\in\mathcal{H}_X, e\in\mathcal{H}_Y}{\operatorname{argmax}} \operatorname{Cov}(\langle X, d\rangle, \langle Y, e\rangle)^2 \quad \text{con } \|d\| = \|e\| = 1. \tag{3.3}$$

Desarrollando la función objetivo del problema de optimización se tiene que

$$\begin{aligned}\operatorname{Cov}(\langle X, d\rangle, \langle Y, e\rangle)^2 &= \mathbb{E}[\langle X, d\rangle\langle Y, e\rangle]^2 \\ &= \mathbb{E}[\langle\langle X, d\rangle Y, e\rangle]^2\end{aligned}$$

$$= \langle e, C_{X,Y}(d)\rangle^2$$

Ahora, consideremos $d \in \mathcal{H}$ fijo. Por la desigualdad de Cauchy-Schwarz tenemos

$$\langle e, C_{X,Y}(d)\rangle^2 \leq \langle e, e\rangle \cdot \langle C_{X,Y}(d), C_{X,Y}(d)\rangle,$$

y la igualdad se alcanza cuando $e \propto C_{X,Y}(d)$, luego, llegamos a un resultado análogo al Lema 2.1.

Lema 3.1

Sea X y Y elementos aleatorios en $L^2_{\mathcal{H}}$. Considerando el problema de optimización (3.3), se cumple que

$$c_1 \propto C_{X,Y}(w_1),$$

y por tanto, (3.3) se puede simplificar a:

$$w_1 = \underset{d \in \mathcal{H}_X}{\operatorname{argmax}} \langle d, C_{Y,X} \circ C_{X,Y}(d)\rangle \quad \textit{sujeto a } \|d\| = 1.$$

Observación 3.3

El Lema 2.1 está descrito en términos de los estimadores muestrales de la matriz de covarianzas, mientras que el Lema 3.1 utiliza los operadores de covarianza teóricos. Sin embargo, el Lema 3.1 es general y también es válido cuando $\mathcal{H}_X$ y $\mathcal{H}_Y$ son de dimensión finita. Por lo tanto, se puede obtener una versión del Lema 2.1 en términos de las matrices de covarianza teóricas.

Observamos que el Lema 3.1 corresponde a la generalización de la primera iteración de PLS, pero utilizando los operadores de covarianza teóricos. Así, dadas las muestras $\{Y_1, \ldots, Y_n\}$ y $\{X_1, \ldots, X_n\}$, sea

$$\hat{C}_{X,Y} = n^{-1} \sum_{i=1}^{n} X_i \otimes Y_i, \tag{3.4}$$

el operador de covarianza muestral. Se define de forma análoga $\hat{C}_{Y,X}$.

Para definir el primer paso de mínimos cuadrados funcionales, basta sustituir los operadores de covarianza teóricos en el Lema 3.1 por sus versiones muestrales.

Posteriormente, se obtienen los estimadores $\hat{w}_1$ y $\hat{c}_1$ a partir del Lema 3.1.

3.3.1. Información residual

Para reiniciar el algoritmo, se debe trabajar con la información residual que no contiene las direcciones estimadas $\hat{w}_1$ y $\hat{c}_1$. Esto implica generalizar al caso funcional el último paso del Algoritmo 1.2.

Además, dado que solo tenemos una variable respuesta, podemos proceder como en el Caso Simple de la Sección 1.1.

Para el caso de dimensión finita, vimos que la información residual correspondiente a la variable explicativa es

$$\mathbf{X} \leftarrow \mathbf{X}\left(I_m - \frac{\hat{w}_1 \hat{w}_1^T \mathbf{X}^T \mathbf{X}}{\hat{w}_1^T \mathbf{X}^T \mathbf{X} \hat{w}_1}\right).$$

Consideramos entonces que cada observación se actualiza como

$$X_i \leftarrow \tilde{P}_{\hat{w}_1} X_i \quad \text{para } i = 1, \ldots, n,$$

con $\tilde{P}_{\hat{w}_1}$ definida en la Sección 2.3.2.

El análogo de la proyección $\tilde{P}_{\hat{w}_1}$ para el caso funcional sería

$$\tilde{P}_{\hat{w}_1} f := f - \frac{\langle f, \hat{C}_X(\hat{w}_1)\rangle}{\langle \hat{w}_1, \hat{C}_X(\hat{w}_1)\rangle} \hat{w}_1 \quad \text{para todo } f \in \mathcal{H}_X, \tag{3.5}$$

donde $\hat{C}_X$ es el operador de covarianza muestral de X.

Por otro lado, definimos los scores

$$T_{1i} := \langle X_i, \hat{w}_1 \rangle \quad \text{para } i = 1, \dots, n.$$

La información residual asociada a la variable Y, se obtiene como en la Sección 1.1.1, se realiza la regresión entre las muestras aleatorias $\{Y_1, \dots, Y_n\}$ y los scores $\{T_{11}, \dots, T_{1n}\}$. Los residuales de esa regresión sería la información residual para Y. Para ver como realizar esa regresión se puede ver el Capítulo 13 de Ramsay y Silverman, 2005.

Por lo tanto, una propuesta de mínimos cuadrados parciales funcionales (FPLS por sus siglas en inglés) es la siguiente

Algoritmo 3.1 (*Algoritmo FPLS*)

Inicialmente se tienen muestras aleatorias $\{Y_1^{(1)}, \dots, Y_n^{(1)}\}$ y $\{X_1^{(1)}, \dots, X_n^{(1)}\}$ de variables en $L^2_{\mathcal{H}_Y}$ y $L^2_{\mathcal{H}_X}$. Además, se tiene un número p fijo de componentes que se quiere calcular.
Para $i = 1, \dots, p$

1. Se obtiene $\hat{w}_i$ resolviendo el problema de optimización en el Lema 3.1 utilizando los operadores $\hat{C}_{X,Y}$ y $\hat{C}_{Y,X}$.

2. Actualizamos la muestra aleatoria asociada a X

$$X_j^{(i+1)} \leftarrow \tilde{P}_{\hat{w}_i} X_j^{(i)} \quad \text{para } j = 1, \dots, n.$$

3. Se obtienen los scores $T_{ij} = \left\langle X_j^{(i)}, \hat{w}_i \right\rangle$. Obtenemos $\{Y_1^{(i+1)}, \dots, Y_n^{(i+1)}\}$ como los residuales de la regresión entre $\{Y_1^{(i)}, \dots, Y_n^{(i)}\}$ y $\{T_{i1}, \dots, T_{in}\}$.

4. Se consideran estas nuevas muestras aleatorias para X y Y y se hace $i \leftarrow i + 1$.

El Algoritmo FPLS es una de las contribuciones innovadoras de este trabajo. Este método se puede usar para sintetizar la información en nuevas variables escalares dentro de un modelo de regresión funcional, simplificando así la complejidad del problema. Cabe destacar que, en principio, la variable respuesta y la variable explicativa toman valores en espacios distintos.

De manera similar a como se hace en el caso multivariado, esta metodología se adaptará para ser aplicada en series de tiempo funcionales.

3.4. FPLS para obtener una proyección estable

3.4.1. Descripción de la metodología

Sea $\{X_n, n = 1, \dots, T\}$ una muestra de una serie de tiempo funcional con media 0 y condición inicial $X_0 = 0$.

De manera similar a (2.2), partimos esta muestra de funciones aleatorias como

$$\mathbf{Y} = \{X_T, \dots, X_2\} \quad \text{y} \quad \mathbf{X} = \{X_{T-1}, \dots, X_1\}. \tag{3.6}$$

En el esquema de regresión inherente al algoritmo PLS, Y representa la variable X_n y X la variable en la serie con un lag anterior, es decir, X_{n-1}.

Utilizando la notación de la sección anterior, observamos que en este caso $\mathcal{H}_X = \mathcal{H}_Y$. Denotemos por $\mathcal{H}$ al espacio de Hilbert común.

En el Algoritmo 2.1, al igual que en cada iteración del Algoritmo FPLS, se requiere actualizar la muestra aleatoria completa $\{X_1, \dots, X_n\}$. Además, la extensión del Algoritmo 2.1 al caso funcional se describe como sigue.

Algoritmo 3.2 (*Algoritmo FPLS para series de tiempo funcionales*)

Comenzamos con una muestra aleatoria $X_n^{(1)}, n = 1, \dots, T$ de una serie de tiempo funcional con valores en $L^2_{\mathcal{H}}$. Sea p un número fijo de componentes a calcular.
Sean **X** y **Y** como en (3.6).
Para $i = 1, \dots, p$

1. Se obtiene $\hat{w}_i$ como en el Lema 3.1 a partir de las muestras aleatorias **X** y **Y**.
2. Se actualiza la muestra aleatoria de la serie como

$$X_n^{(i+1)} \leftarrow \tilde{P}_{\hat{w}_i} X_n^{(i)} \quad \text{para } n = 1, \dots, T,$$

   con $\tilde{P}_{\hat{w}_i}$ definido en (3.5).
3. Se forman **X** y **Y** como en (3.6) con la muestra aleatoria obtenida en el paso 2, y luego hacemos $i \leftarrow i + 1$.

Observación 3.4

Un resultado que puede investigarse es el análogo del Lema 2.2. Si consideramos que $\hat{w}_1$ es asintóticamente equivalente al primer componente principal de la serie X, entonces

$$\tilde{P}_{\hat{w}_1} f \approx f - \langle \hat{w}_1, f \rangle \hat{w}_1 \quad \text{para toda } f \in \mathcal{H},$$

donde $\tilde{P}_{\hat{w}_1}$ es la proyección sobre $\text{SPAN}\{\hat{w}_1\}^{\perp}$. Por lo tanto, el Algoritmo 3.2 es asintóticamente equivalente al algoritmo que utiliza FPCA, descrito en la Sección 3.2.1.

Durante la ejecución del Algoritmo 3.2, se reduce simultáneamente la varianza del proceso utilizando las componentes $\hat{w}_i$ que poseen capacidad predictiva. Según Seo, 2023a, las primeras direcciones $\hat{w}_i$ pueden resultar en proyecciones no estacionarias, es decir, las series de scores

$$T_i(n) := \left\langle X_n^{(i)}, \hat{w}_i \right\rangle \quad \text{para } n = 1, \dots, T,$$

pueden ser no estacionarias.

Esta observación, similar a la discutida en la Sección 2.3.3, permitirá estimar una base para $\mathcal{H}^N$ y consecuentemente estimar φ, bajo el Supuesto 3.2.

3.4.2. Estimación de la proyección estable

Tanto en el método descrito en la Sección 3.2.1 como en el Algoritmo 3.2, es necesario identificar el valor de φ.

Al igual que en la Sección 2.3.3, se utilizarán los scores

$$T_i(n) = \left\langle X_n^{(i)}, \hat{w}_i \right\rangle \quad \text{para } n = 1, \dots, T,$$

para cada $\hat{w}_i$ obtenido mediante el Algoritmo 3.2 o el i-ésimo componente principal si se utiliza FPCA.

Esperamos que para todo $i \leq \varphi$, la muestra T_i corresponda a una serie de tiempo no estacionaria, lo que motiva la siguiente definición.

Definición 3.4 (*Espacio Inestable estimado*)

Sea $\{X_n, n = 1, \dots, T\}$ una muestra de una serie de tiempo funcional con valores en $L^2_{\mathcal{H}}$.
Sea

$$\hat{\varphi} = \begin{cases} \text{máx}\{j \mid T_j \text{ es no estacionaria}\} & \text{si } T_1 \text{ es no estacionaria,} \\ 0 & \text{de otra forma.} \end{cases}$$

Definimos el espacio inestable estimado por

$$\mathcal{H}^N = \begin{cases} \text{SPAN}\{\hat{w}_1, \dots, \hat{w}_{\hat{\varphi}}\} & \text{si } \hat{\varphi} \neq 0, \\ \{0\} & \text{en otro caso.} \end{cases}$$

Por lo tanto, se itera el Algoritmo 3.2 o se obtienen componentes principales hasta que se obtenga por primera vez una serie de scores estacionaria; esta serie debería ser la serie $T_{\hat{\varphi}+1}$.

Si utilizamos el Algoritmo 3.2 para definir la base $\{\hat{w}_1, \dots, \hat{w}_{\hat{\varphi}}\}$, los estimadores de P^N y P^S son:

$$\hat{P}^N_{\text{FPLS}} := \sum_{j=1}^{\hat{\varphi}} \hat{w}_j \otimes \hat{w}_j \quad \text{y} \quad \hat{P}^S_{\text{FPLS}} := I - \hat{P}^N_{\text{FPLS}}.$$

Si se utiliza FPCA, los estimadores se denotan como en (3.2), sustituyendo φ por su estimador $\hat{\varphi}$.

Para evaluar la estacionaridad de T_i, se puede emplear una prueba KPSS. Si esta prueba es rechazada, existe evidencia de que T_i no es estacionario. Por lo tanto, el estimador $\hat{\varphi}$ se obtiene cuando por primera vez se rechaza la hipótesis nula de que $T_{\hat{\varphi}+1}$ es una serie estacionaria.

Por supuesto, se puede utilizar otra prueba de hipótesis para determinar la base del espacio inestable, y esta condicionará los resultados. Así, la metodología empleada puede denominarse en función del método utilizado para obtener las componentes $\hat{w}_i$ (FPLS o FPCA) y la prueba de hipótesis utilizada para determinar la base del espacio inestable estimado (KPSS, ADF, etc.).

Así, el esquema de hipótesis sobre el valor de φ, que a su vez determina la base de $\mathcal{H}^N$, es:

$$H_0: \underbrace{\varphi = \varphi_0}_{T_{\varphi_0+1} \text{ es I(0)}} \quad \text{vs} \quad H_1: \underbrace{\varphi > \varphi_0}_{T_{\varphi_0+1} \text{ es I(1)}} .$$

Observamos que para definir la base del espacio inestable se puede analizar la estacionaridad de $\{P^S_{\varphi_0} X_n, n = 1, \dots, T\}$ para cada φ_0, donde $P^S_{\varphi_0}$ es la proyección sobre $\text{SPAN}\{\hat{w}_1, \dots, \hat{w}_{\varphi_0}\}^\perp$. Por lo tanto, se establece un esquema de prueba de hipótesis como sigue:

$$H_0: \underbrace{\varphi = \varphi_0}_{P^S_{\varphi_0} X \text{ es estacionaria}} \quad \text{vs} \quad H_1: \underbrace{\varphi > \varphi_0}_{P^S_{\varphi_0} X \text{ no es estacionaria}} .$$

Se podría utilizar, por ejemplo, la prueba descrita en Horváth et al., 2014 para analizar la estacionaridad de $\{P^S_{\varphi_0} X_n, n = 1, \dots, T\}$.

La implementación computacional de este trabajo consideró al menos las pruebas KPSS y ADF para los scores, así como la prueba descrita en Horváth et al., 2014. Posteriormente, se analizará cuál de estas pruebas tiene un mejor desempeño respecto al esquema de simulaciones propuesto.

3.5. Implementación Computacional

El objetivo de esta sección es detallar la implementación computacional del Algoritmo 3.2. En primer lugar, es necesario considerar algunos aspectos teóricos que simplifiquen principalmente el problema de optimización en el Algoritmo 3.2.

3.5.1. Aspectos teóricos

En la práctica, comúnmente se considera que el espacio de Hilbert donde toman valores las variables es $\mathcal{H} = L^2[0,1]$. En adelante, al mencionar $\mathcal{H}$ nos referiremos a este espacio de Hilbert específico.

En primer lugar, debemos implementar el problema de optimización

$$w_1 = \underset{d \in \mathcal{H}_X}{\text{argmax}} \langle d, C_{Y,X} \circ C_{X,Y}(d) \rangle^2 \quad \text{con} \quad \|d\| = 1, \tag{3.7}$$

como se describe en el Lema 3.1.

Consideremos para el resto de la sección

$$\Psi := C_{Y,X} \circ C_{X,Y},$$

donde, como en el Algoritmo 3.2, X y Y son variables en $L^2_{\mathcal{H}}$. Con estas consideraciones, se enuncia la siguiente proposición.

Proposición 3.3

El operador Ψ es simétrico, positivo y de Hilbert-Schmidt.

Demostración. En primer lugar, veamos que Ψ es simétrico.

Notamos, por propiedades del operador de covarianza cruzada, que

$$\begin{aligned}\Psi^\star &= (C_{Y,X} \circ C_{X,Y})^\star \\ &= (C_{X,Y})^\star \circ (C_{Y,X})^\star \\ &= C_{Y,X} \circ C_{X,Y} \\ &= \Psi.\end{aligned}$$

Así, Ψ es simétrico o auto adjunto. Ahora veamos que es positivo, para ello basta recordar nuevamente que $C^\star_{Y,X} = C_{X,Y}$, entonces, para todo $f \in \mathcal{H}$

$$\langle f, \Psi f \rangle = \langle C_{X,Y} f, C_{X,Y} f \rangle \geq 0,$$

por propiedades del producto interior, luego, Ψ es positivo.

Finalmente, notamos que $C_{X,Y}$ (y $C_{Y,X}$) es son operadores nucleares (Ver Bosq, 2000, Identidad 1.62), en particular, son de Hilbert-Schmidth. Debido a que el conjunto de operadores de Hilbert-Schmidt $\mathcal{S}$ es un ideal (Ver Dummit & Foote, 2004, pp. 242) se tiene que $\Psi \in \mathcal{S}$, que es lo que queríamos probar. □

Por lo tanto, según el Teorema 1.4, el problema de optimización (3.7) se reduce a encontrar la eigenfunción asociada al eigenvalor más grande del operador Ψ. En el caso del Algoritmo 3.2, se utiliza la versión muestral

$$\hat{\Psi} := \hat{C}_{Y,X} \circ \hat{C}_{X,Y}.$$

En el Análisis de Datos Funcionales, no se suele estimar directamente, por ejemplo, $\hat{C}_{X,Y}$ como en (3.4); en cambio, si es posible, se demuestra que es un operador integral y lo que se estima es su kernel. Este es el caso, por ejemplo, para el operador de covarianza, como se ejemplifica en la Proposición 1.9.

Proposición 3.4

El operador $\Psi := C_{Y,X} \circ C_{X,Y}$ es un operador integral con kernel ψ dado por

$$\psi(s,t) := \int_0^1 c(u,t)c(u,s)du,$$

donde c es la función de covarianza entre Y y X.

Demostración. Notamos que

$$\begin{aligned}\Psi(x)(t) &= C_{Y,X} \circ C_{X,Y}(x)(t) \\ &= \mathbb{E}\big[\langle C_{X,Y}(x), Y\rangle X_t\big].\end{aligned}$$

Para $\omega \in \Omega$ fijo se tiene que

$$\langle C_{X,Y}(x), Y(\omega)\rangle = \int_0^1 C_{X,Y}(x)(u) Y_u(\omega) du.$$

Así

$$\Psi(x)(t) = \int_0^1 \mathbb{E}[X_t Y_u] C_{X,Y}(x)(u) du.$$

Sea $c(u,t) := \mathbb{E}[Y_u X_t]$ entonces

$$\Psi(x)(t) = \int_0^1 c(u,t) C_{X,Y}(x)(u) du.$$

Usando el Teorema de Fubini, tenemos

$$\begin{aligned} C_{X,Y}(x)(u) &= \mathbb{E}[\langle x, X\rangle Y_u] \\ &= \int_\Omega \int_0^1 x(s) X_s(\omega) Y_u(\omega) ds d\mathbb{P}(\omega) \\ &= \int_0^1 \mathbb{E}[X_s Y_u] x(s) ds \\ &= \int_0^1 c(u,s) x(s) ds. \end{aligned}$$

Por lo tanto

$$\Psi(x)(t) = \int_0^1 \int_0^1 c(u,t) c(u,s) x(s) ds du.$$

Definimos

$$\psi(s,t) := \int_0^1 c(u,t) c(u,s) du. \tag{3.8}$$

Nuevamente, por el Teorema de Fubini, tenemos que

$$\Psi(x)(t) = \int_0^1 \psi(s,t) x(s) ds, \tag{3.9}$$

que es lo que queríamos probar. □

Este resultado es muy importante para simplificar los cálculos en los estimadores muestrales, que se definen en la siguiente sección.

3.5.2. Estimadores muestrales

Las ecuaciones (3.8) y (3.9) representan las expresiones teóricas del operador Ψ y su kernel ψ. A partir de estas expresiones, se construyen los estimadores muestrales. Nos enfocaremos en la implementación para series de tiempo funcionales.

En el contexto del Algoritmo 3.2, Y representa la variable X_n y X corresponde a la variable X_{n-1}. Es decir, para la muestra aleatoria $\{X_n, n = 1, \ldots, T\}$ de la serie de tiempo funcional, el estimador $\hat{c}$ de la función de covarianza que aparece en la Proposición 3.4 es

$$\hat{c}(u,v) = (T-1)^{-1} \sum_{i=2}^{T} X_i(u) X_{i-1}(v).$$

En teoría, todo elemento $f \in \mathcal{H}$ puede expresarse en términos de una base ortonormal numerable (ver Lang, 2012, pp. 102), siempre y cuando $\mathcal{H}$ sea separable.

Así, para una base ortonormal numerable $\{B_n, n \in \mathbb{N}\}$ del espacio de Hilbert $\mathcal{H}$, se tiene:

$$f = \sum_{j=1}^{\infty} \langle f, B_j \rangle B_j$$

En la práctica, como se mencionó anteriormente, se tiene que $\mathcal{H} = L^2[0,1]$, y las bases comúnmente utilizadas son la base de Fourier o bases de B-Splines (De Boor, 1986).

Computacionalmente, dada la base $\{B_n, n \in \mathbb{N}\}$, se necesita truncar la expresión en serie. Esto, por supuesto, induce una pérdida de información. Sin embargo, esta aproximación está justificada, ya que:

$$\lim_{n\to\infty} \langle f, B_n \rangle = 0.$$

Consideremos un m suficientemente grande de manera que truncamos la serie hasta tener m sumandos. Sea $\{B_1, \ldots, B_m\}$ los elementos básicos correspondientes, que acomodaremos en un vector columna:

$$B(t) = [B_1(t), \ldots, B_m(t)]^T.$$

Así,

$$X_n(t) = \mathbf{c}_n^T B(t) \quad \text{para } t \in [0,1] \text{ y } n = 1, \ldots, T,$$

donde $\mathbf{c}_n \in \mathbb{R}^m$ representa los coeficientes respecto a la base $\{B_1, \ldots, B_m\}$ de la función X_n para $n = 1, \ldots, T$. De hecho, $(\mathbf{c}_n)_k = \langle X_n, B_k \rangle$ para $k = 1, \ldots, m$.

Respecto a ese sistema de base, el estimador muestral $\hat{c}$ es

$$\hat{c}(u,v) = (T-1)^{-1} \cdot B(u)^T \mathbf{C}_{2:T} \mathbf{C}_{1:T-1}^T B(v), \tag{3.10}$$

donde

$$\mathbf{C}_{2:T} = [\mathbf{c}_2, \ldots, \mathbf{c}_T] \in \mathbb{R}^{m\times(T-1)} \quad \text{y} \quad \mathbf{C}_{1:T-1} = [\mathbf{c}_1, \ldots, \mathbf{c}_{T-1}] \in \mathbb{R}^{m\times(T-1)}.$$

Observación 3.5

La igualdad en (3.10) en realidad es una aproximación debido al truncamiento de la base. En lo sucesivo, se escribirán siempre igualdades, aunque debe tenerse en mente que son, en realidad, aproximaciones.

En este notación, el estimador muestral $\hat{\psi}$ viene dado por

$$\begin{aligned}
\hat{\psi}(s,t) &= \int_0^1 \hat{c}(u,t)\hat{c}(u,s)du \\
&= (T-1)^{-2} \int_0^1 \big(B(u)^T \mathbf{C}_{2:T} \mathbf{C}_{1:T-1}^T B(t)\big) \cdot \big(B(u)^T \mathbf{C}_{2:T} \mathbf{C}_{1:T-1}^T B(s)\big) du. \\
&= (T-1)^{-2} \cdot B(t)^T \mathbf{C}_{1:T-1} \mathbf{C}_{2:T}^T \cdot \left(\int_0^1 B(u) B(u)^T du \right) \cdot \mathbf{C}_{2:T} \mathbf{C}_{1:T-1}^T B(s).
\end{aligned}$$

Observamos que podemos expresar a $x \in \mathcal{H}$ en términos de la base B de la siguiente manera:

$$x(s) = \mathbf{c}_x^T B(s),$$

donde el subíndice indica que los coeficientes dependen del funcional x.

Así

$$\begin{aligned}
\hat{\Psi}(x)(t) &= \int_0^1 \hat{\psi}(s,t) x(s) ds \\
&= (T-1)^{-2} \cdot \int_0^1 B(t)^T \mathbf{C}_{1:T-1} \mathbf{C}_{2:T}^T \cdot \left(\int_0^1 B(u) B(u)^T du \right) \cdot \mathbf{C}_{2:T} \mathbf{C}_{1:T-1}^T B(s) B(s)^T \mathbf{c}_x ds.
\end{aligned}$$

Sea $\mathbf{B} = \int_0^1 B(s)B(s)^T ds$ que es una matriz simétrica de $m \times m$. De hecho, $\mathbf{B} = I_m$, con I_m la matriz identidad de $m \times m$, pues la base que estamos considerando es ortonormal.

Por lo tanto,

$$\hat{\Psi}(x)(t) = (T-1)^{-2} \cdot B(t)^T \cdot \mathbf{C}_{1:T-1} \mathbf{C}_{2:T}^T \mathbf{C}_{2:T} \mathbf{C}_{1:T-1}^T \cdot \mathbf{c}_x.$$

Sea

$$\mathbf{A}_{\hat{\Psi}} := \mathbf{C}_{1:T-1} \mathbf{C}_{2:T}^T \mathbf{C}_{2:T} \mathbf{C}_{1:T-1}^T \in \mathbb{R}^{m\times m},$$

por lo que

$$\hat{\Psi}(x)(t) = (T-1)^{-2} B(t)^T \mathbf{A}_{\hat{\Psi}} \mathbf{c}_x.$$

Si $\mathbf{v}$ es eigenvector de $\mathbf{A}_{\hat{\Psi}}$ entonces

$$\mathbf{A}_{\hat{\Psi}}\mathbf{v} = \lambda \cdot \mathbf{v},$$

para algún $\lambda \in \mathbb{R}^+$, ya que $\mathbf{A}_{\hat{\Psi}}$ es positiva y simétrica.

Sea $v(t) = \mathbf{v}^T B(t)$ para todo $t \in [0,1]$. Considerando lo anterior se tiene

$$\begin{aligned}\hat{\Psi}(v)(t) &= (T-1)^{-2} B(t)^T (\lambda \cdot \mathbf{v}) \\ &= (T-1)^{-2} \cdot \lambda v(t).\end{aligned}$$

Concluimos que v es aproximadamente una eigenfunción del operador $\hat{\Psi}$.

Observación 3.6

Resolver el problema de eigenvalores para $\hat{\Psi}$ es aproximadamente equivalente a resolver el problema de eigenvalores para $\mathbf{A}_{\hat{\Psi}}$.

En el contexto del algoritmo FPLS, si $\mathbf{c}_{\hat{w}_1}$ es el coeficiente asociado al valor propio más grande de $\mathbf{A}_{\hat{\Psi}}$, la eigenfunción correspondiente, $\hat{w}_1$, que resuelve el problema de optimización en el Lema 3.1, es

$$\hat{w}_1(t) = \mathbf{c}_{\hat{w}_1}^T B(t) \quad \text{para } t \in [0,1],$$

con $\|\mathbf{c}_{\hat{w}_1}\| = 1$, lo que implica que $\|\hat{w}_1\| = 1$.

Esta simplificación del problema de optimización hace que la implementación computacional sea muy similar a la del caso multivariado. En particular, el Algoritmo 3.2 se reduce a aplicar el algoritmo PLS sobre los coeficientes del proceso, respecto a una base ortonormal truncada de $\mathcal{H}$.

3.5.3. Información residual

Aquí detallaremos cómo se puede aproximar la proyección $\tilde{P}_{\hat{w}_1}$, definida en (3.5), para actualizar el proceso $\{X_n, n = 1, \dots, T\}$.

La idea es utilizar la base truncada $\{B_1, \dots, B_m\}$ para realizar todos los cálculos, de manera que todo se reduzca al caso multivariado, como se describió en la sección anterior.

Observamos que

$$\begin{aligned}\langle X_n, \hat{w}_1 \rangle &= \left\langle \mathbf{c}_n^T B, \mathbf{c}_{\hat{w}_1}^T B \right\rangle \\ &= \sum_{i=1}^{m}\sum_{j=1}^{m} \left\langle (\mathbf{c}_n)_i B_i), (\mathbf{c}_{\hat{w}_1})_j B_j \right\rangle \\ &= \sum_{i=1}^{m}\sum_{j=1}^{m} (\mathbf{c}_n)_i (\mathbf{c}_{\hat{w}_1})_j \left\langle B_i, B_j \right\rangle.\end{aligned}$$

Como $\{B_1, \dots, B_m\}$ es una base ortonormal entonces $\left\langle B_i, B_j \right\rangle = \delta_{ij}$ donde δ_{ij} es la delta de Kronecker, por lo que

$$\langle X_n, \hat{w}_1 \rangle = \mathbf{c}_n^T \mathbf{c}_{\hat{w}_1}.$$

De hecho, acabamos de demostrar que, para aproximar el producto interior de dos elementos funcionales, es suficiente con calcular el producto interior entre los coeficientes asociados a la base truncada.

Aplicando el mismo razonamiento se tiene que

$$\langle X_n, \hat{C}_X(\hat{w}_1) \rangle = \mathbf{c}_n^T \mathbf{C}_{1:T} \mathbf{C}_{1:T}^T \mathbf{c}_{\hat{w}_1},$$

y

$$\langle \hat{w}_1, \hat{C}_X(\hat{w}_1) \rangle = \mathbf{c}_{\hat{w}_1}^T \mathbf{C}_{1:T} \mathbf{C}_{1:T}^T \mathbf{c}_{\hat{w}_1},$$

Por lo tanto, la información residual en el Algoritmo 3.2, en términos de la base truncada de B, es

$$\tilde{P}_{\hat{w}_1} X_n = \mathbf{c}_n^T \left(I_m - \frac{\mathbf{C}_{1:T} \mathbf{C}_{1:T}^T \mathbf{c}_{\hat{w}_1} \mathbf{c}_{\hat{w}_1}^T}{\mathbf{c}_{\hat{w}_1}^T \mathbf{C}_{1:T} \mathbf{C}_{1:T}^T \mathbf{c}_{\hat{w}_1}} \right) B(t).$$

Esta es la proyección que se realizará en cada paso del Algoritmo 3.2 para obtener la información residual.

Hemos reducido todos los cálculos de cada uno de los pasos principales del Algoritmo 3.2 a operaciones matriciales, tomando como referencia la base $\{B_1, \dots, B_m\} \subset \mathcal{H}$.

3.5.4. Proyección al espacio estable

Con las direcciones calculadas $\hat{w}_1, \dots, \hat{w}_\varphi$, procederemos a definir la proyección al espacio $\mathcal{H}_N$ de tendencias comunes y, por consiguiente, al espacio estable $\mathcal{H}_S$.

El supuesto $\varphi := \dim \mathcal{H}_N < \infty$ es muy importante, pues nos permite obtener la estimación de P^N, que denotaremos por $\hat{P}^N$, sin truncamiento. El estimador $\hat{P}^N$ viene dado por:

$$\begin{aligned}\hat{P}^N X_n(t) &= \sum_{i=1}^{\varphi} \langle X_n, \hat{w}_i \rangle \hat{w}_i \\ &= \sum_{i=1}^{\varphi} (\mathbf{c}_n^T \mathbf{c}_{\hat{w}_i}) \cdot \mathbf{c}_{\hat{w}_i}^T B(t),\end{aligned}$$

por lo que la estimación de la proyección al espacio estable, $\hat{P}^S$, es

$$\begin{aligned}\hat{P}^S X_n(t) &= X_n(t) - \hat{P}^N X_n(t) \\ &= \mathbf{c}_n^T B(t) - \sum_{i=1}^{\varphi} (\mathbf{c}_n^T \mathbf{c}_{\hat{w}_i}) \cdot \mathbf{c}_{\hat{w}_i}^T B(t) \\ &= \mathbf{c}_n^T \cdot \left(I_m - \sum_{i=1}^{\varphi} \mathbf{c}_{\hat{w}_i} \mathbf{c}_{\hat{w}_i}^T \right) \cdot B(t).\end{aligned}$$

De nueva cuenta, en notación matricial se tiene

$$\hat{P}^S \mathbf{X}(t) = \mathbf{C}_{1:T}^T \cdot \left(I_m - \sum_{i=1}^{\varphi} \mathbf{c}_{\hat{w}_i} \mathbf{c}_{\hat{w}_i}^T \right) \cdot B(t),$$

donde

$$\mathbf{X}(t) := [X_1(t), \dots, X_n(t)]^T.$$

Así, la matriz de coeficientes con respecto a la base $\{B_1, \dots, B_m\}$ de la proyección estable de la serie funcional es

$$\left(I_m - \sum_{i=1}^{\varphi} \mathbf{c}_{\hat{w}_i} \mathbf{c}_{\hat{w}_i}^T \right) \cdot \mathbf{C}_{1:T}.$$

Consideremos la matriz de coeficientes anterior, dado que es el formato requerido por la librería fda (Ramsay, 2023). Las columnas de la matriz representan los coeficientes, asociados a una base determinada, de cada observación.

3.6. Esquema de predicción

En el caso multivariado, se pudieron calcular todos los componentes $\hat{w}_1, \dots, \hat{w}_m$, y algunos de ellos correspondían a una base para el espacio estable $\mathcal{H}_S$. Los scores T_i asociados a las componentes estables representan la información conjunta de las variables de la serie de tiempo, actuando como un indicador. Estos scores permiten interpretar, en función de las cargas, el efecto de cada variable.

Sin embargo, en el caso funcional, bajo el Supuesto 3.2, el espacio estable $\mathcal{H}^S$ es de dimensión infinita. Además, al estimar $\hat{P}^S$, ya sea mediante FPCA o FPLS, en ningún momento estimamos las eigenfunciones asociadas a $\mathcal{H}^S$.

En cambio, se estima la proyección

$$\eta_n := \hat{P}^S X_n \quad \text{para } n = 1, 2, \dots, T.$$

Podemos realizar una regresión análoga a la propuesta en la Sección 2.3.4 para validar cuánta información se perdió con la proyección al espacio estable.

Eso es simplemente un proceso de validación y exploración que permite visualizar el grado de confianza en la proyección al espacio estable estimada.

La forma propuesta para realizar esta validación es a través de un esquema de regresión funcional. Se aplicará el *modelo de concurrencia* descrito en la Sección 14.1 de Ramsay y Silverman, 2005.

Este modelo en resumen supone una relación de la forma

$$X_n(t) = \alpha(t)\eta_n(t) + \epsilon_n(t),$$

donde $\{\epsilon_n, n = 1, 2, \ldots, T\}$ es una muestra de un ruido blanco fuerte en $L^2_{\mathcal{H}}$.

Para realizar el modelo de regresión anterior se utiliza la función fRegress de la librería fda (Ramsay, 2023). En el caso funcional, al momento de proporcionar una base para la aproximación mediante su truncamiento, es necesario penalizar la regresión para obtener un coeficiente α suave. Esto se hace como se describe en el Capítulo 5 de Ramsay et al., 2009. Para cada una de las proyecciones, ya sea con FPLS o FPCA, se utiliza el mismo parámetro de suavizamiento para garantizar una comparación justa.

Además, se pueden generar pronósticos con un error controlado del proceso η, el cual proviene por construcción de una serie de tiempo funcional estacionaria. La metodología para este caso está presentada en Aue et al., 2015. Este paso podría considerarse análogo al uso del modelo VAR para predecir el proceso de scores estables.

En general, cuando se utiliza FPCA, es relevante la interpretación proporcionada por las eigenfunciones que concentran la mayor varianza y sus scores asociados. Podemos realizar un análisis similar con las primeras componentes estables obtenidas por FPLS, ya que además de concentrar la variabilidad explicada y su estabilidad, tienen capacidad predictiva.

De esta manera, se define un esquema predictivo asociado a la proyección al espacio estable estimado (Algoritmo 2.2).

Algoritmo 3.3

Iniciamos con una muestra aleatoria $\{X_n, n = 1, \ldots, T\}$ y un horizonte de predicción h.

1. Obtenemos el estimador $\hat{P}^S$ de la proyección al espacio estable $\mathcal{H}^S$, lo cual se expresa como

$$\eta_n := \hat{P}^S X_n.$$

2. Obtenemos predicciones $\hat{\eta}_{T+1}, \ldots, \hat{\eta}_{T+h}$ usando la metodología de Aue et al., 2015.

El ajuste realizado al proceso η y sus predicciones se pueden considerar como los pronósticos del proceso X en el espacio estable $\mathcal{H}^S$.

Tras realizar la predicción de la proyección estacionaria, se pueden generar componentes estables con sus scores correspondientes. Así, mediante el Algoritmo 3.2, se pueden obtener eigenfunciones $\hat{s}_1, \ldots, \hat{s}_p$ con sus correspondientes scores

$$S_i(n) := \langle \eta_n, \hat{s}_i \rangle \quad \text{para } n = 1, \ldots, T,$$

que podemos interpretar, ya que concentran información predictiva de la parte estable de la serie de tiempo funcional original.

Concluimos que se puede realizar un análisis sobre la parte estable de X, que es η, para obtener información sobre el proceso original.

3.7. Resultados y Conclusiones

3.7.1. Simulaciones

Utilizando las proyecciones teóricas al espacio estable y a su complemento ortogonal, definimos

$$Z_n^N := P^N \Delta X_n, \quad Z_n^S := P^S \eta_n \quad \text{para } n \geq 0,$$

y

$$Z_n := Z_n^N + Z_n^S \quad \text{para } n \geq 0.$$

En el caso de cointegración, η está definido en la Descomposición de Beveridge-Nelson.

Observación 3.7

Bajo el Supuesto 3.1, la Descomposición de Beveridge-Nelson implica que

$$Z_n^S = P^S X_n \quad \text{para } n \geq 0.$$

Esto refuerza la idea de que se puede entender a η como la proyección de X al espacio $\mathcal{H}^S$.

Los procesos Z^N y Z^S nos permitirán simular el proceso X para un valor de φ dado. Los detalles sobre cómo se implementará esto se presentan en la siguiente sección.

Utilizaremos el esquema de simulación descrito en detalle en Seo, 2023b. Recordemos que $\varphi = \dim \mathcal{H}^N$. Consideremos $\{B_j, j \in \mathbb{N}\}$ una base de Fourier en $\mathcal{H} = L^2[0,1]$.

Como se discute en Seo, 2023b, para valores de φ muy grandes, la precisión en la estimación de las proyecciones disminuye. Por esta razón, al igual que en Seo, 2023b, consideraremos $\varphi = 2, 3, 4, 5$.

Así, se construyen los procesos Z^N y Z^S de la siguiente forma:

$$Z_n^N = \sum_{j=1}^{\varphi} \alpha_j \langle B_j^N, Z_{n-1} \rangle B_j^N + P^N \epsilon_n \quad \text{y} \quad Z_n^S = \sum_{j=1}^{10} \beta_j \langle B_j^S, Z_{n-1} \rangle B_j^S + P^S \epsilon_n \quad \text{para } n = 1, \ldots, T,$$

donde

$$\{B_j^N, j = 1, \ldots, \varphi\} \subset \{B_1, \ldots, B_6\} \quad \text{(muestreo sin reemplazo)}$$
$$\{B_j^S, j = 1, \ldots, 10\} \subset \{B_7, \ldots, B_{18}\} \quad \text{(muestreo sin reemplazo).}$$

Observación 3.8

Se considera que $Z_0 = 0$, para evitar lidiar con el problema del efecto de la condición inicial. De otra forma, para cada simulación se debería ignorar una cantidad considerable de las primeras iteraciones. Esta idea también se aplica al caso multivariado.

El ruido blanco ϵ se construye como

$$\epsilon_n = \sum_{j=1}^{80} \theta_{j,n} (0.95)^{j-1} B_j,$$

donde $\{\theta_{j,n}\}$ es una realización de una familia de variables aleatorias normales estándar independientes en j y n. No obstante, también se puede considerar una distribución distinta, por ejemplo, una t-Student, para evaluar el efecto de las metodologías sobre errores con colas pesadas.

Construimos X a partir de Z notando que

$$X_n = \sum_{j=1}^{n} P^N \Delta X_j + P^S X_n + P^N X_0 = \sum_{j=1}^{n} Z_j^N + Z_n^S + P^N X_0.$$

Simulamos $\{\alpha_j, j = 1, \ldots, \varphi\}$ y $\{\beta_j, j = 1, \ldots, 10\}$ como

$$\alpha_j \sim \mathcal{U}[-0.5, 0.5] \quad \text{y} \quad \beta_j \sim \mathcal{U}[\beta_{\min}, \beta_{\max}],$$

donde $\mathcal{U}(a,b)$ representa la distribución uniforme en el intervalo $[a,b]$.

Los coeficientes β tienen la interpretación de ser el grado de *persistencia* de la no estacionaridad de la serie Z^S. Como la estacionaridad es un concepto asintótico, valores más pequeños de β indican que la serie alcanza la estacionaridad más rápidamente.

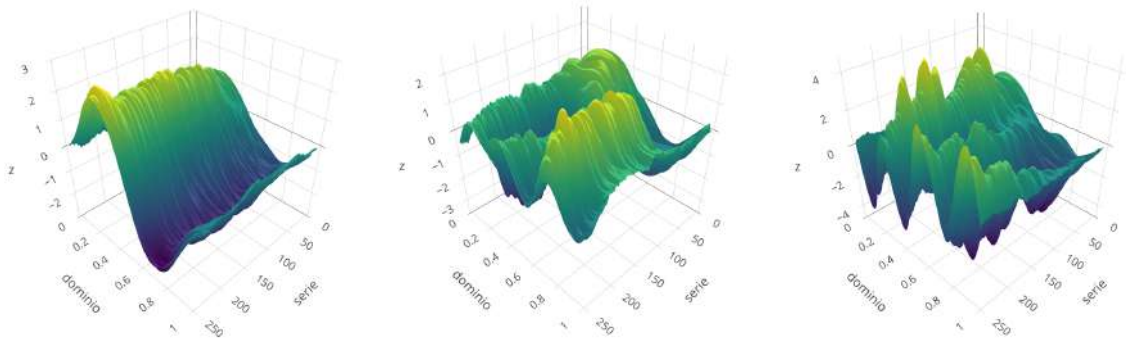
La persistencia tiene un efecto en las pruebas de hipótesis para identificar el espacio estable, se analizan los siguientes casos:

1. Baja persistencia: $[\beta_{\text{min}}, \beta_{\text{max}}] = [0, 0.5]$
2. Alta persistencia: $[\beta_{\text{min}}, \beta_{\text{max}}] = [0.5, 0.7]$

Se utilizan todos estos elementos para calcular los valores de X_n en 200 puntos equidistantes en $[0,1]$. Luego, se suaviza X_n mediante una regresión penalizada, tal como se describe en el Capítulo 5 de Ramsay et al., 2009.

Para visualizar una serie temporal funcional, se utiliza una representación tridimensional. Por ejemplo, el eje de las ordenadas representa el valor de n. El eje de las abscisas corresponde al dominio de las funciones, que en este caso es $[0,1]$. El eje Z, por su parte, representa la imagen de X_n.

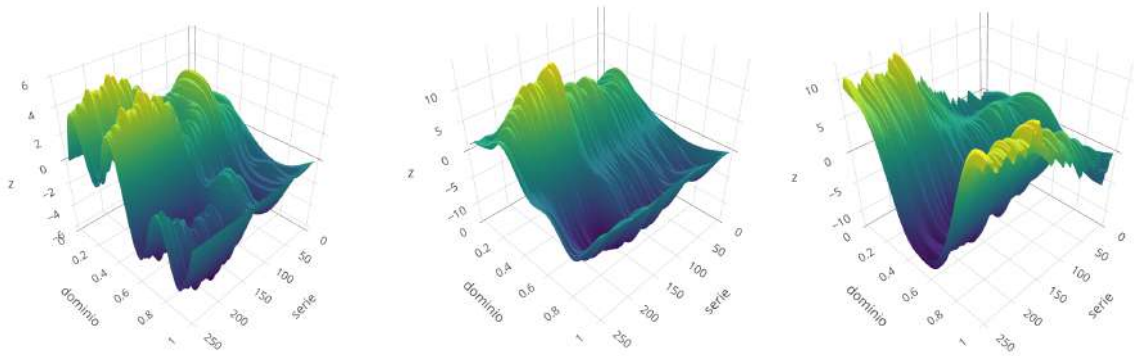
Visualizamos los resultados de simular series de longitud $T = 250$. En las Figuras 6 y 7 se muestran los resultados para $\varphi = 3, 4, 5$, considerando baja o alta persistencia. Se puede observar que con alta persistencia



(a) (b) (c)

Figura 6.: Diferentes simulaciones del proceso X considerando en todas ellas un esquema de baja persistencia. En (a), se tiene $\varphi = 3$; en (b), $\varphi = 4$; y en (c), el espacio estable es tal que $\varphi = 5$.

el retorno a la media es más lento que con baja persistencia.



(a) (b) (c)

Figura 7.: Diferentes simulaciones del proceso X considerando en todas ellas un esquema de alta persistencia. En (a), se tiene $\varphi = 3$; en (b), $\varphi = 4$; y en (c), el espacio estable es tal que $\varphi = 5$.

3.7.2. Resultados

A continuación, describiremos la simulación que ilustrará la metodología de validación y predicción.

Simularemos una serie con condición inicial $X_0 = 0$ y longitud $T = 100$. El ruido seguirá una distribución t-Student con 3 grados de libertad. La serie simulada tendrá $\varphi = 4$.

Para esta ilustración, utilizaremos un esquema de alta persistencia. La simulación específica para ejemplificar la metodología se muestra en la Figura 8.

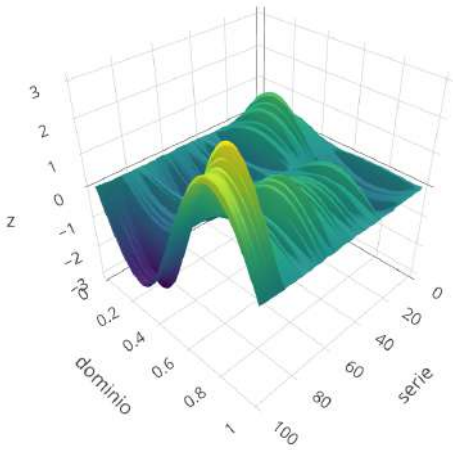
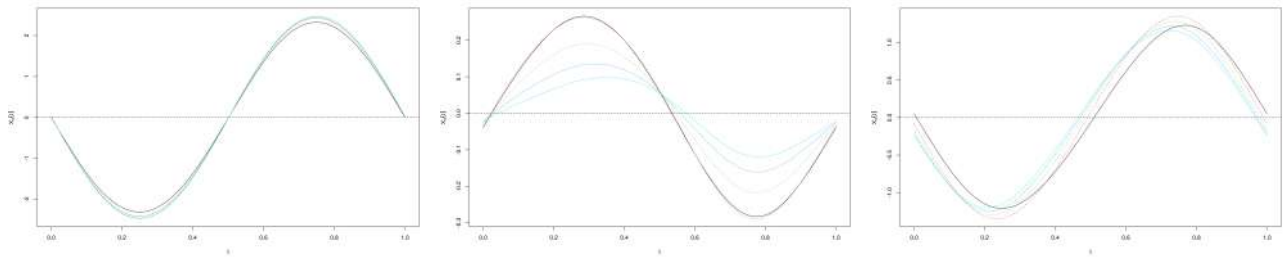


Figura 8.: Simulación de una serie de tiempo funcional con longitud $T = 100$ y $\varphi = 4$.

Validaremos el grado de pérdida de información de la serie de tiempo funcional después de aplicar el Algoritmo 3.3, utilizando tanto FPLS como FPCA. Además, emplearemos la prueba ADF para identificar los componentes básicos de $\mathcal{H}^N$, que empíricamente mostró ser la más efectiva para estimar φ. En ambas metodologías se obtuvo $\hat{\varphi} = 4$.

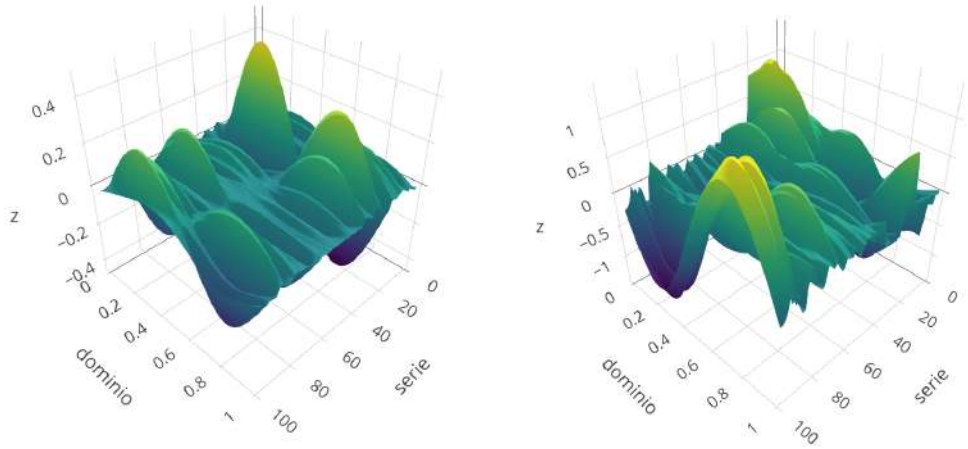
Utilizaremos un horizonte de predicción $h = 5$, es decir, estimaremos la proyección estable con $\{X_n, n = 1, \dots, T - h\}$. Luego, calcularemos $\eta_1, \dots, \eta_{T-h}$ y los pronósticos $\eta_{T-h+1}, \dots, \eta_T$ mediante el método descrito en Aue et al., 2015.

Con $\{\eta_n, n = 1, \dots, T-h\}$ y $\{X_n, n = 1, \dots, T-h\}$ se ajusta el modelo de concurrencia. Mediante este modelo se estiman valores para la serie. Los resultados se presentan en las Figuras 10a y 10b. La razón de considerar el horizonte de predicción es visualizar en qué medida se recupera información que inicialmente no se utilizó para la estimación de la proyección estable.



(a) Valores observados (b) FPCA (c) FPLS

Figura 9.: Valores recuperados mediante el modelo de concurrencia con las predicciones $\eta_{T-h+1}, \ldots, \eta_T$ y $h = 5$.



(a) FPCA (b) FPLS

Figura 10.: Se presentan los resultados del modelo de concurrencia al utilizar FPLS y FPCA para obtener la proyección $\hat{P}^S$.

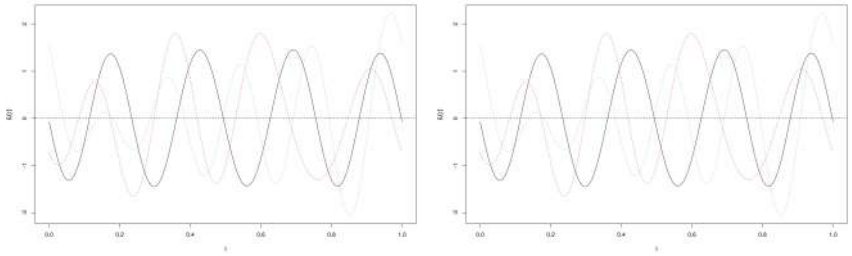
En la Figura 9, se observa cómo se recuperaron las últimas h funciones. Se nota que los valores observados $X_{T-h+1}, \ldots, X_T$ se recuperan mejor con la proyección estimada por FPLS que por FPCA. Esto proporciona una mayor fiabilidad a los scores y eigenfunciones estimadas por FPLS.

Visualmente notamos que FPLS recupera mejor la forma de la superficie de la serie de tiempo funcional simulada. Además, calculamos las eigenfunciones $\hat{s}_i$ y los scores asociados S_i de η. Estas primeras eigenfunciones son componentes estables representativas de la serie original.

En la Figura 11, observamos que la primera componente es muy similar en ambos métodos, pero hay diferencias en la segunda y tercera componente.

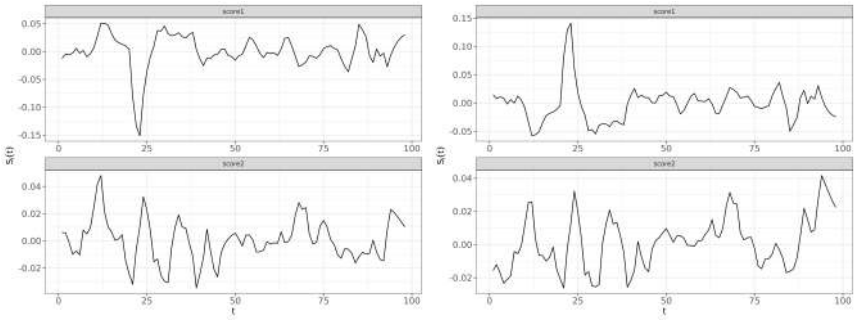
Estas diferencias son más evidentes al observar los scores correspondientes a las componentes. Podemos comparar los valores de S_1 y S_2 en cada método. Cada uno de los scores corresponde a series de tiempo univariadas. En la Figura 12, se observa que el primer score para cada método parece tener principalmente diferencias en el signo, lo cual es relevante para su interpretación, como se discutió en la Sección 2.4. En contraste, el segundo score muestra diferencias más sustanciales entre los métodos.

Por lo tanto, podemos inferir información valiosa sobre la serie de tiempo original a través de la dinámica de sus scores estables, que podemos predecir como se describe en la Sección 2.4, ya sea mediante un modelo VAR o individualmente. Esto último es factible debido a la ortogonalidad de las componentes



(a) FPCA (b) FPLS

Figura 11.: Se muestran las primeras tres componentes $\hat{s}_i$ de η para cada método. La primera componente se representa con línea sólida negra, la segunda con línea roja y la tercera con línea verde.



(a) FPCA (b) FPLS

Figura 12.: Se muestra la gráfica de los scores S_1 y S_2 para cada método, visualizados como series de tiempo.

estimadas.

Debido a que FPLS recupera mejor la información que FPCA, tanto en la estimación como en la predicción fuera de muestra, los scores asociados a FPLS ofrecen información más fiable. La dinámica de estos scores resulta útil para entender el comportamiento futuro de la serie de tiempo funcional.

3.8. Conclusiones y Trabajo Futuro

La metodología de mínimos cuadrados parciales (PLS), tanto en su formulación de dimensión finita como en su versión de dimensión infinita (caso funcional), representa una alternativa flexible y de fácil implementación para sintetizar información en scores estacionarios que contengan la mayor cantidad de información posible. Esto aborda de manera efectiva el problema común de no estacionaridad y alta dimensionalidad en el análisis de series de tiempo.

En este trabajo de tesis, se presenta una contribución significativa mediante la generalización del algoritmo PLS para el caso funcional. En un contexto de regresión, esta generalización permite relacionar dos elementos aleatorios funcionales, ofreciendo una perspectiva única que difiere de las propuestas presentadas en Delaigle y Hall, 2012 o Preda y Saporta, 2005. Se ha adaptado esta generalización para su aplicación en series de tiempo funcionales, de manera similar a su aplicación en el caso multivariado.

En el caso analizado, se aborda la ausencia de tendencia determinística, mientras que se discute cómo manejar esta situación en el contexto multivariado. Sin embargo, en el caso funcional, la estimación de tales tendencias es un área de investigación actual. Un trabajo relevante en este sentido es Martínez-Hernández y Genton, 2021. Como posible dirección futura, se propone explorar el uso del algoritmo FPLS para abordar automáticamente las tendencias determinísticas, y también considerar una versión modificada similar a la propuesta en Seo, 2023a.

Además, se destaca la conexión tanto en el caso funcional como en el de dimensión finita con el concepto de cointegración. Para futuras investigaciones, se plantea la necesidad de formalizar la noción de espacio estable introducida, con el fin de definir estadísticamente la proyección estimada y obtener resultados de consistencia, como el Teorema 3.1 de Seo, 2023a. Se espera que este concepto a formalizar sea más general que la noción de cointegración, aunque en el contexto funcional, esta última sigue siendo un área de investigación en evolución.

En esa misma línea de investigación, sería interesante explorar la posibilidad de demostrar un resultado análogo al Lema 2.2 para el caso funcional, en el contexto de cointegración funcional descrito en el Supuesto 3.1.

Además, dado que en cada iteración, tanto el algoritmo PLS como el FPLS buscan direcciones que maximicen la covarianza, en realidad están buscando funcionales lineales que maximicen la dependencia lineal entre las variables. En este sentido, sería viable adaptar la función objetivo a optimizar, ya sea en (2.3) o (3.3), dependiendo de si se trata del caso multivariado o funcional, para maximizar otro tipo de dependencias. Esta dependencia podría ser modelada utilizando la Teoría de Cópulas (Czado, 2019; Nelsen, 2006), lo que permitiría identificar componentes que capturen otro tipo de dependencia en la serie y que, al mismo tiempo, sean estacionarias. Esta perspectiva representa una línea de investigación futura que podría enriquecer significativamente la metodología propuesta.

El objetivo principal de este trabajo es enfatizar la importancia de proponer metodologías flexibles y no paramétricas que aborden eficazmente el desafío de la no estacionaridad en series de tiempo. De este modo, se pretende analizar los fenómenos subyacentes integrando el rigor matemático con los aspectos fundamentales de la modelación y la implementación computacional. Además, se lleva a cabo un estudio original del algoritmo PLS, con la aspiración de consolidarlo como un método de referencia junto a sus similares PCA y CCA.

§ A. Apéndice del Capítulo 1

Prueba Proposición 1.1.

1. Supongamos que $i < j$. Notamos que

$$\mathbf{V}_j = \left(I_n - \frac{T_{j-1} T_{j-1}^T}{T_{j-1}^T T_{j-1}} \right) \mathbf{V}_{j-1}.$$

Definamos $Z_k = I_n - \frac{T_k T_k^T}{T_k^T T_k}$. Recursivamente, llegamos a que

$$\mathbf{V}_j = Z_{j-1} \cdots Z_i \mathbf{V}_i.$$

Ahora, notamos que

$$Z_i \mathbf{V}_i \hat{w}_i = \left(I_n - \frac{T_i T_i^T}{T_i^T T_i} \right) T_i = T_i - T_i = 0,$$

lo que implica que

$$\mathbf{V}_j \hat{w}_i = 0 \quad \text{para } j > i.$$

Por otro lado, por el Algoritmo 1.2, sabemos que

$$\hat{w}_j \propto \mathbf{V}_j^T \mathbf{u}_j.$$

Por lo tanto,

$$\hat{w}_j^T \hat{w}_i \propto \mathbf{u}_j^T \mathbf{V}_j \hat{w}_i = 0,$$

así que $\hat{w}_j^T \hat{w}_i = 0$ para todo $j > i$, lo que implica que $\{\hat{w}_1, \ldots, \hat{w}_m\}$ son mutuamente ortogonales.

2. De manera análoga, sea $j > i$. En este caso, notamos que

$$\mathbf{V}_j = \mathbf{V}_{j-1} \left(I_m - \frac{\hat{w}_{j-1} \hat{w}_{j-1}^T \mathbf{V}_{j-1}^T \mathbf{V}_{j-1}}{\hat{w}_{j-1}^T \mathbf{V}_{j-1}^T \mathbf{V}_{j-1} \hat{w}_{j-1}} \right).$$

Consideremos $Z_k = I_m - \frac{\hat{w}_k \hat{w}_k^T \mathbf{V}_k^T \mathbf{V}_k}{\hat{w}_k^T \mathbf{V}_k^T \mathbf{V}_k \hat{w}_k}$. De esta forma, llegamos a que

$$\mathbf{V}_j = \mathbf{V}_i Z_i \cdots Z_{j-1}.$$

Observamos que

$$T_i^T \mathbf{V}_i Z_i = \hat{w}_i^T \mathbf{V}_i^T \mathbf{V}_i \left(I_m - \frac{\hat{w}_i \hat{w}_i^T \mathbf{V}_i^T \mathbf{V}_i}{\hat{w}_i^T \mathbf{V}_i^T \mathbf{V}_i \hat{w}_i} \right) = \hat{w}_i^T \mathbf{V}_i^T \mathbf{V}_i - \hat{w}_i^T \mathbf{V}_i^T \mathbf{V}_i = 0.$$

Esto implica que $T_i^T \mathbf{V}_j = 0$, y a su vez

$$T_i^T T_j = T_i^T \mathbf{V}_j \hat{w}_j = 0 \quad \text{para } j > i,$$

por lo que $\{T_1, \ldots, T_m\}$ son mutuamente ortogonales. □

Prueba Proposición 1.2. Sabemos por el Algoritmo 1.1 que $\hat{c}_i$ es un eigenvector asociado al eigenvalor más grande $\lambda > 0$ de $\mathbf{R}_i^T \mathbf{V}_i \mathbf{V}_i^T \mathbf{R}_i$.

Usando la equivalencia de los algoritmos 1.2 y 1.1, tenemos:

$$\hat{w}_i = k \cdot \mathbf{V}_i^T \mathbf{R}_i \hat{c}_i, \tag{A.1}$$

para alguna constante k. De hecho, como $\|\hat{w}_i\| = \|\hat{c}_i\| = 1$, entonces:

$$\begin{aligned} 1 &= k^2 \cdot \hat{c}_i^T \mathbf{R}_i^T \mathbf{V}_i \mathbf{V}_i^T \mathbf{R}_i \hat{c}_i \\ &= k^2 \lambda \hat{c}_i^T \hat{c}_i \\ &= k^2 \lambda, \end{aligned}$$

luego, $k = \lambda^{-\frac{1}{2}}$.

Multiplicando por $\mathbf{R}_i^T \mathbf{V}_i$ a (A.1) y usando que $\hat{c}_i$ es eigenvector, tenemos:

$$\hat{c}_i = \lambda^{-\frac{1}{2}} \mathbf{R}_i^T \mathbf{V}_i \hat{w}_i,$$

y sustituyendo en (A.1), llegamos a que:

$$\hat{w}_i = \lambda^{-1} \mathbf{V}_i^T \mathbf{R}_i \mathbf{R}_i^T \mathbf{V}_i \hat{w}_i,$$

por lo que $\hat{w}_i$ es un eigenvector asociado al valor más grande de $\mathbf{V}_i^T \mathbf{R}_i \mathbf{R}_i^T \mathbf{V}_i$, ya que el espectro de $\mathbf{V}_i^T \mathbf{R}_i \mathbf{R}_i^T \mathbf{V}_i$ y de $\mathbf{R}_i^T \mathbf{V}_i \mathbf{V}_i^T \mathbf{R}_i$ es el mismo.

Finalmente, consideremos la descomposición en valores singulares de la matriz de covarianza muestral $\mathbf{V}_i^T \mathbf{R}_i$ (Ver Giraud, 2021, Teorema C.1):

$$\mathbf{V}_i^T \mathbf{R}_i = \sum_{j=1}^{r} \lambda_j \mathbf{f}_j \mathbf{g}_j^T,$$

con $r = \operatorname{rk} \mathbf{V}_i^T \mathbf{R}_i$, $\lambda_1 \geq \cdots \geq \lambda_r > 0$ los eigenvalores de $\mathbf{V}_i^T \mathbf{R}_i \mathbf{R}_i^T \mathbf{V}_i$ (y de $\mathbf{R}_i^T \mathbf{V}_i \mathbf{V}_i^T \mathbf{R}_i$), y $\{\mathbf{f}_1, \ldots, \mathbf{f}_r\}$ (resp. $\{\mathbf{g}_1, \ldots, \mathbf{g}_r\}$) los eigenvectores asociados de $\mathbf{V}_i^T \mathbf{R}_i \mathbf{R}_i^T, \mathbf{V}_i$ (resp. $\mathbf{R}_i^T, \mathbf{V}_i \mathbf{V}_i^T \mathbf{R}_i$).

Por lo tanto, $\lambda = \lambda_1$, $\hat{w}_i = \mathbf{f}_1$ y $\hat{c}_i = \mathbf{g}_1$.

Finalmente, notamos que:

$$\widehat{\operatorname{Cov}}(\mathbf{V}_i d, \mathbf{R}_i e)^2 \propto (d^T \mathbf{V}_i^T \mathbf{R}_i e)^2.$$

Por la desigualdad de Cauchy-Schwarz:

$$\widehat{\operatorname{Cov}}(\mathbf{V}_i d, \mathbf{R}_i e)^2 \leq \langle \mathbf{R}_i^T \mathbf{V}_i d, \mathbf{R}_i \mathbf{V}_i^T d \rangle \cdot \langle e, e \rangle = \langle \mathbf{R}_i^T \mathbf{V}_i d, \mathbf{R}_i \mathbf{V}_i^T d \rangle,$$

pues $\|e\| = 1$ y donde $\langle \cdot, \cdot \rangle$ representa el producto interior en el espacio euclideano correspondiente.

Por lo tanto, si d es fijo, el máximo anterior se alcanza cuando $e \propto \mathbf{R}_i \mathbf{V}_i^T d$.

Así, considerando $e \propto \mathbf{R}_i \mathbf{V}_i^T d$, entonces:

$$\widehat{\operatorname{Cov}}(\mathbf{V}_i d, \mathbf{R}_i e)^2 \propto d^T \mathbf{V}_i^T \mathbf{R}_i \mathbf{R}_i^T \mathbf{V}_i d.$$

El valor máximo de la identidad anterior se alcanza para $d = \hat{w}_i$, lo que implica $e = \hat{c}_i$, lo que se quería probar. □

§ B. Apéndice del Capítulo 2

Prueba Teorema 2.1. Consideremos la siguiente notación:

$$A := \beta(\beta^T\beta)^{-1},$$
$$Q^{-1} := [A, \beta_\perp],$$
$$\Gamma(z) := I_m - \sum_{i=1}^{p-1} \Gamma_i z^i.$$

Con ello, definimos:

$$B^\star(z) := Q[\Gamma(z)A(1-z) - \alpha z, \Gamma(z)\beta_\perp],$$
$$B(z) := Q^{-1}B^\star(z)Q.$$

En primer lugar, si definimos:

$$P(z) := \begin{bmatrix} \beta^\perp \\ (1-z)A_\perp^T \end{bmatrix},$$

notamos que:

$$P(z) = \begin{bmatrix} I_r & 0 \\ 0 & I_{m-r} \end{bmatrix} Q. \tag{B.1}$$

Observamos que:

$$C(z) = Q^{-1}B^\star(z)P(z).$$

Por la propiedad multiplicativa del determinante, tenemos que:

$$\det C(z) = \det Q^{-1} \cdot \det B^\star(z) \cdot \det P(z).$$

Por la ecuación (B.1), se tiene que el factor $\det P(z)$ es el que comprende las $m-r$ raíces unitarias de $\det C(z)$, por lo que:

$$\det B^\star(z) \neq 0 \quad \text{para} \quad |z| \leq 1. \tag{B.2}$$

Ahora, definimos:

$$Z_n := Q^{-1}P(L)X_n.$$

Observamos que:

$$Z_n = A\beta^T X_n + \beta_\perp A_\perp^T \Delta X_n.$$

Notamos que $A_\perp = \beta_\perp(\beta_\perp^T\beta_\perp)^{-1}$. Luego, si P_β es la matriz de proyección a Col β (resp. $P_{\beta_\perp}$ la proyección a Col $\beta_\perp$), entonces:

$$P_\beta = A\beta^T \quad \text{y} \quad P_{\beta_\perp} = \beta_\perp A_\perp^T,$$

y así:

(B.3) $$Z_n = P_\beta X_n + P_{\beta_\perp}\Delta X_n.$$

Además, se tiene que:

$$B(L)Z_n = C(L)X_n = \epsilon_n,$$

y observamos que:

$$B(0) = [A, \beta_\perp]Q = I_m,$$

luego $B(z)$ tiene la forma:

$$B(z) = I_m - \sum_{i=1}^{p} B_i z^i.$$

Como comentario adicional, de esto se sigue que Z es un modelo VAR(p) estable.
Además, de (B.2), se sigue que Z es invertible (Ver Brockwell & Davis, 1991, Definición 3.1.4). Luego, si:

$$\theta(z) := B(z)^{-1} = \sum_{j=0}^{\infty} \theta_j z^j,$$

entonces:

(B.4) $$Z_n = \theta(L)\epsilon_n.$$

Será útil parametrizar $\theta(z)$ como:

$$\theta(z) = \theta(1) + (1-z)\theta^\star(z),$$

con $\theta^\star(z) := \sum_{j=0}^{\infty} \theta_j^\star z^j$. Comparando coeficientes de ambas expresiones, se puede deducir que:

$$\begin{aligned}
\theta_0^\star &= \theta_0 - \theta(1), \\
\theta_i^\star &= -\sum_{j=i+1}^{\infty} \theta_j \quad \text{para} \quad i \geq 1.
\end{aligned}$$

Con esta parametrización, obtenemos que:

$$Z_n = \theta(1)\epsilon_n + \theta^\star(L)\Delta\epsilon_n.$$

Por otro lado, por (B.3), se tiene $P_{\beta_\perp} Z_n = P_{\beta_\perp}\Delta X_n$. Luego:

$$\begin{aligned}
X_n &= X_0 + \sum_{i=1}^{n} \Delta X_i, \\
&= X_0 + \sum_{i=1}^{n} \left(P_\beta \Delta X_i + P_{\beta_\perp}\Delta X_i \right), \\
&= X_0 + \sum_{i=1}^{n} P_{\beta_\perp} Z_i + \sum_{i=1}^{n} P_\beta \Delta X_i.
\end{aligned}$$

Sustituyendo la parametrización de la descomposición MA de Z, tenemos:

$$X_n = X_0 + P_\beta X_n + P_{\beta_\perp}\theta(1)\sum_{i=1}^{n}\epsilon_i + P_{\beta_\perp}\theta^\star(L)(\epsilon_n - \epsilon_0) - P_\beta X_0.$$

Sea:

$$X_0^\star := X_0 - P_\beta X_0 - P_{\beta_\perp}\theta^\star(L)\epsilon_0.$$

Así,

$$X_n = X_0^\star + P_{\beta_\perp}\left(\theta(1)\sum_{i=1}^{n}\epsilon_i\right) + P_{\beta_\perp}\theta^\star(L)\epsilon_n + P_\beta X_n. \tag{B.5}$$

Para la parte estacionaria, de (B.3) y (B.4), se sigue que:

$$P_\beta X_n = P_\beta Z_n = \Big(P_\beta\theta(L)\Big)\epsilon_n.$$

Luego, si definimos:

$$\xi_n := P_{\beta_\perp}\theta^\star(L)\epsilon_n + P_\beta X_n,$$

entonces el proceso ξ es estacionario con representación MA:

$$\xi_n = \Big(P_{\beta_\perp}\theta^\star(L) + P_\beta\theta(L)\Big)\epsilon_n.$$

Precisamente, definimos:

$$\mathfrak{F}^\star(z) := P_{\beta_\perp}\theta^\star(z) + P_\beta\theta(z),$$

luego $\xi_n = \mathfrak{F}^\star(L)\epsilon_n$.

Para la parte de caminatas aleatorias o tendencias, notamos:

$$P_{\beta_\perp}\theta(1) = P_{\beta_\perp}Q^{-1}B^\star(1)^{-1}Q,$$

$$= P_{\beta_\perp}[A,\beta_\perp][-\alpha,\Gamma(1)\beta_\perp]^{-1}.$$

En Lütkepohl, 2005, se nota que:

$$[-\alpha,\Gamma(1)\beta_\perp]^{-1} = \begin{bmatrix}(\alpha^T\alpha)^{-1}\alpha^T(\Gamma(1)\mathfrak{F} - I_m)\\ \left(\alpha_\perp^T\Gamma(1)\beta_\perp\right)^{-1}\alpha_\perp^T\end{bmatrix}.$$

Haciendo todas las cuentas y considerando esta identidad, llegamos a que:

$$P_{\beta_\perp}\theta(1) = \mathfrak{F}.$$

Concluimos de (B.5) que:

$$X_n = X_0^\star + \mathfrak{F}\cdot\sum_{i=1}^{n}\epsilon_i + \mathfrak{F}^\star(L)\epsilon_n,$$

que es lo que queríamos probar. □

Prueba Proposición 2.1. Para demostrar (2.1), basta con observar la identidad (B.5). Definimos $\eta_n := P_{\beta_\perp}\theta^\star(L)\epsilon_n$, que sabemos es un proceso estacionario, y tenemos la expresión:

$$X_n = X_0^\star + P_{\beta_\perp}\left(\theta(1)\sum_{i=1}^{n}\epsilon_i\right) + P_\beta X_n + \eta_n,$$

que es precisamente (2.1).

El proceso Z que aparece en la Proposición 2.1 se define en la prueba del Teorema 2.1, y la identidad (B.3) es la descomposición presentada en la proposición. La representación MA es la obtenida en la identidad (B.4). □

Prueba Lema 2.2. Sean $\mathbf{X}$ y $\mathbf{Y}$ como en (2.2), y consideremos los estimadores de las covarianzas asintóticas:

$$S_{01} := T^{-1}\mathbf{X}^T\mathbf{Y} = T^{-1}\sum_{i=2}^{T} X_{i-1}X_i^T,$$

y

$$S_{10} := T^{-1}\mathbf{Y}^T\mathbf{X} = T^{-1}\sum_{i=2}^{T} X_i X_{i-1}^T.$$

Por el Lema 2.1, se tiene que $\hat{w}_1$ y $\hat{c}_1$ resuelven los problemas de eigenvalores:

$$S_{01}S_{10}\hat{w}_1 = \lambda\hat{w}_1,$$
$$S_{10}S_{01}\hat{c}_1 = \lambda\hat{c}_1,$$

donde $S_{10} = S_{01}^T$ y λ corresponde al valor singular más grande de S_{01}.

Consideremos las normalizaciones:

$$S_{01} := \tilde{T}^{-\frac{3}{2}}S_{01}$$

y

$$S_{10} := S_{01}^T.$$

Por hipótesis, sabemos que X cumple las condiciones del Teorema de Representación de Granger, por lo que

$$X_n = N_n + \eta_n + X_0^\star,$$

con $N \sim I(1)$, $\eta \sim I(0)$ y $X_0^\star$ una constante.

Como buscamos un resultado asintótico, podemos suponer sin perder generalidad que $X_0^\star = 0$.

De lo anterior, tenemos que

$$S_{01} = T^{-\frac{3}{2}}\left(\sum_{i=2}^{T} N_{i-1}N_i^T + \sum_{i=2}^{T} N_{i-1}\eta_i^T + \sum_{i=2}^{T}\eta_{i-1}N_i^T + \sum_{i=2}^{T}\eta_{i-1}\eta_i^T\right). \tag{B.6}$$

Por el Teorema 2.1, se tiene que

$$\eta_n = \mathfrak{F}^\star(B)\epsilon_n = \sum_{i=0}^{\infty}\mathfrak{F}_i^\star\epsilon_{n-i} \quad \text{y} \quad N_n = \mathfrak{F}\cdot\sum_{i=1}^{n}\epsilon_i.$$

Así, por el Teorema B.13 de Johansen, 1995, se cumple, primero, que

$$T^{-1}\sum_{i=2}^{T} N_{i-1}\eta_i^T \quad \Longrightarrow \quad \mathfrak{F}\cdot\int_0^1 W_i(dW_i)^T\cdot\mathfrak{F}^\star(1)^T + \sum_{i=1}^{\infty}\mathfrak{F}_i^\star,$$

donde $\Longrightarrow$ representa convergencia débil de procesos y W es un movimiento browniano m-dimensional (Ver Karatzas & Shreve, 2014, Definición 5.1).

Para este término, utilizando el Teorema de Slutsky y recordando que la convergencia débil es equivalente a la convergencia en probabilidad cuando el límite es una constante, tenemos que

$$T^{-\frac{3}{2}}\sum_{i=2}^{T} N_{i-1}\eta_i^T \xrightarrow{P} 0 \quad \Leftrightarrow \quad T^{-\frac{3}{2}}\sum_{i=2}^{T} N_{i-1}\eta_i^T = o_P(1).$$

Por otro lado, ya que η es un proceso estacionario con representación MA bien definida, se tiene que

$$T^{-1}\sum_{i=2}^{T}\eta_{i-1}\eta_i^T \xrightarrow{P} \Gamma_\eta(1),$$

por lo que nuevamente se concluye

$$T^{-\frac{3}{2}} \sum_{i=2}^{T} \eta_{i-1}\eta_i^T = o_P(1).$$

Ahora, notamos que

$$\begin{aligned}\sum_{i=2}^{T} \eta_{i-1}N_i^T &= \left(\sum_{i=2}^{T} N_i\eta_{i-1}^T\right)^T \\ &= \left(\sum_{i=2}^{T}\left(\sum_{j=1}^{i-2}\epsilon_j\right)\eta_{i-1}^T + \epsilon_{i-1}\eta_{i-1}^T + \epsilon_i\eta_{i-1}^T\right)^T \cdot \mathfrak{F}^T \\ &= \left(\sum_{i=2}^{T}\left(\sum_{j=1}^{i-2}\epsilon_j\right)\eta_{i-1}^T\right)^T \cdot \mathfrak{F}^T + \left(\sum_{i=2}^{T}\epsilon_{i-1}\eta_{i-1}^T\right)^T \cdot \mathfrak{F}^T + \left(\sum_{i=2}^{T}\epsilon_i\eta_{i-1}^T\right)^T \cdot \mathfrak{F}^T.\end{aligned}$$

Utilizando nuevamente el Teorema B.13 de Johansen, 1995, concluimos que los tres sumandos del lado derecho de la identidad anterior multiplicados por $T^{-\frac{3}{2}}$ también son variables $o_P(1)$, es decir,

$$T^{-\frac{3}{2}} \sum_{i=2}^{T} \eta_{i-1}N_i^T = o_P(1).$$

De todo lo anterior y (B.6), concluimos que

$$S_{01} = T^{-\frac{3}{2}} \sum_{i=2}^{T} N_{i-1}N_i^T + o_p(1).$$

De forma análoga, podemos probar que

$$S_{10} = T^{-\frac{3}{2}} \sum_{i=1}^{T} N_i N_{i-1}^T + o_P(1).$$

Finalmente, observamos que

$$T^{-\frac{3}{2}} \sum_{i=1}^{T} N_i N_i^T = T^{-\frac{3}{2}} \sum_{i=2}^{T} N_{i-1}N_i^T + T^{-\frac{3}{2}} \cdot \mathfrak{F} \sum_{i=1}^{T} \epsilon_i N_i^T.$$

Otra vez, por el Teorema B.13 de Johansen, 1995, se sigue que

$$T^{-\frac{3}{2}} \cdot \mathfrak{F} \sum_{i=1}^{T} \epsilon_i N_i^T = o_p(1).$$

De esta forma, se tiene que

$$S_{01} = S + o_p(1) \quad \text{y} \quad S_{10} = S + o_p(1).$$

Por lo tanto, $\hat{w}_1$ es asintóticamente un eigenvector de la matriz S^2. Al ser S una matriz simétrica, si tiene descomposición espectral,

$$S = UDU^T,$$

con U matriz unitaria de eigenvectores, entonces

$$S^2 = UD^2U^T.$$

Por lo tanto, los eigenvectores de S^2 son también los de S. Así, $\hat{w}_1$ es asintóticamente el eigenvector asociado al eigenvalor más grande de S, que es lo que queríamos probar. □

§ Bibliografía

Aue, A., Norinho, D. D., & Hörmann, S. (2015). On the prediction of stationary functional time series. *Journal of the American Statistical Association*, *110*(509), 378-392.

Beare, B. K., Seo, J., & Seo, W.-K. (2017). Cointegrated linear processes in Hilbert space. *Journal of Time Series Analysis*, *38*(6), 1010-1027.

Beare, B. K., & Seo, W.-K. (2020). Representation of I (1) and I (2) autoregressive Hilbertian processes. *Econometric Theory*, *36*(5), 773-802.

Billingsley, P. (2013). *Convergence of probability measures*. John Wiley & Sons.

Bosq, D. (2000). *Linear process in function spaces: theory and applications*. Springer-Verlag.

Brockwell, P. J., & Davis, R. A. (1991). *Time Series: Theory and Methods*. Springer.

Chang, Y., Kim, C. S., & Park, J. Y. (2016). Nonstationarity in time series of state densities. *Journal of Econometrics*, *192*(1), 152-167.

Christensen, R., et al. (2011). *Plane answers to complex questions* (4.ª ed.). Springer.

Conway, J. B. (2019). *A course in functional analysis* (Vol. 96). Springer.

Czado, C. (2019). Analyzing dependent data with vine copulas. *Lecture Notes in Statistics, Springer*, *222*.

De Boor, C. (1986). *B (asic)-spline basics*. University of Wisconsin-Madison. Mathematics Research Center.

Delaigle, A., & Hall, P. (2012). Methodology and Theory for Partial Least Squares Applied to Functional Data. *The Annals of Statistics*, *40*, 322-352. https://doi.org/10.1214/11-AOS958

Dummit, D. S., & Foote, R. M. (2004). *Abstract algebra* (Vol. 3). Wiley Hoboken.

Engel, R., & Granger, C. W. J. (1987). Co-Integration and Error Correction: Representation, Estimation and Testing. *Econometrica*, *55*, 251-276.

Escoufier, Y. (1970). Echantillonage dans une population de variables aléatoires réelles. *Publications de l'Institut de Statistique de l'Université de Paris*, *19*, 1-47.

Garthwaite, P. H. (1994). An Interpretation of Partial Least Squares.

Giraud, C. (2021). *Introduction to high-dimensional statistics*. Chapman; Hall/CRC.

Gonzalo, J., & Lee, T.-H. (2000). On the robustness of cointegration tests when series are fractionally intergrated. *Journal of Applied Statistics*, *27*(7), 821-827.

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

Harris, D. (1997). Principal components analysis of cointegrated time series. *Econometric Theory*, *13*(4), 529-557.

Horváth, L., & Kokoszka, P. (2012). *Inference for Functional Data with Applications*. Springer-Verlag.

Horváth, L., Kokoszka, P., & Rice, G. (2014). Testing stationarity of functional time series. *Journal of Econometrics*, *179*(1), 66-82.

Höskuldsson, A. (1988). PLS regression methods. *Journal of Chemometrics*, *2*, 211-228. https://doi.org/10.1002/cem.1180020306

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of educational psychology*, *24*(6), 417.

Hotelling, H. (1936). Relations between two sets of variables. *Biometrika*, *28*(3), 321-377.

Hyndman, R., Athanasopoulos, G., Bergmeir, C., Caceres, G., Chhay, L., O'Hara-Wild, M., Petropoulos, F., Razbash, S., Wang, E., & Yasmeen, F. (2023). *forecast: Forecasting functions for time series and linear models* [R package version 8.21.1]. https://pkg.robjhyndman.com/forecast/

Hyndman, R., & Khandakar, Y. (2008). Automatic Time Series Forecasting: The forecast Package for R. *Journal of Statistical Software*, *27*(3), 1-22. https://doi.org/10.18637/jss.v027.i03

Hyndman, R., Koehler, A., Ord, J. K., & Snyder, R. (2008). *Forecasting with Exponencial Smoothing*. Springer-Verlag.

Johansen, S. (1995). *Likelihood-Based Inference in Cointegrated Vector Auto-Regressive Models*. Oxford University Press.

Johansen, S. (1991). Estimation and hypothesis testing of cointegration vectors in Gaussian vector autoregressive models. *Econometrica: journal of the Econometric Society*, 1551-1580.

Johnson, R. A., Wichern, D. W., et al. (2002). *Applied multivariate statistical analysis*. Prentice hall Upper Saddle River, NJ.

Karatzas, I., & Shreve, S. (2014). *Brownian motion and stochastic calculus* (Vol. 113). springer.

Kwiatkowski, D., Phillips, P., Schmidt, P., & Shin, Y. (1992). Testing the null hypothesis of stationarity against the alternative of a unit root. *Journal of Econometrics*, *54*, 159-178.

Lang, S. (2012). *Real and functional analysis* (Vol. 142). Springer Science & Business Media.

Lütkepohl, H. (2005). *New Introduction to Multiple Time Series Analysis*. Springer-Verlag.

Maddala, G. S., & Kim, I.-M. (1998). Unit roots, cointegration, and structural change.

Martinez Hernandez, I., & Genton, M. G. (2019). Generalized Records for Functional Time Series with Application to Unit Root Tests.

Martínez-Hernández, I., & Genton, M. G. (2021). Nonparametric trend estimation in functional time series with application to annual mortality rates. *Biometrics*, *77*(3), 866-878.

Martínez-Hernández, I., Gonzalo, J., & González-Farías, G. (2022). Nonparametric estimation of functional dynamic factor model. *Journal of Nonparametric Statistics*, *34*(4), 895-916.

Mateos-Aparicio, G. (2011). Partial Least Squares (PLS) Methods: Origins, Evolution, and Application to Social Sciences. *Communications in Statistics - Theory and Methods*, *40*(13), 2305-2317. https://doi.org/10.1080/03610921003778225

Muriel, N., González-Farias, G., & Ramos, R. (2012). A PLS Approach to Cointegration Analysis. *Journal of Business & Economics*, *4*, 177-199.

Nelsen, R. B. (2006). *An introduction to copulas*. Springer.

Pfaff, B. (2008). *Analysis of Integrated and Cointegrated Time Series with R* (Second) [ISBN 0-387-27960-1]. Springer. https://www.pfaffikus.de

Preda, C., & Saporta, G. (2005). PLS regression on a stochastic process [Partial Least Squares]. *Computational Statistics & Data Analysis*, *48*(1), 149-158. https://doi.org/https://doi.org/10.1016/j.csda.2003.10.003

Preda, C., Saporta, G., & Lévéder, C. (2007). PLS Classification of Functional Data. *Computational Statistics*, *22*, 223-235.

Preda, C., & Schiltz, J. (2011). Functional PLS regression with functional response: the basis expansion approach. *Proceedings of the 14th Applied Stochastic Models and Data Analysis Conference*, 1126-1133.

Ramsay, J., Hooker, G., & Graves, S. (2009). *Functional Data Analysis with R and MATLAB*. Springer-Verlag.

Ramsay, J. (2023). *fda: Functional Data Analysis* [R package version 6.1.4]. https://CRAN.R-project.org/package=fda

Ramsay, J., & Silverman, B. (2005). *Functional Data Analysis*. Springer Verlag.

Saxe, K. (2002). *Beginning functional analysis*. Springer.

Seo, W.-K. (2023a). Functional Principal Component Analysis for Cointegrated Functional Time Series. *Journal of Time Series Analysis*. https://doi.org/10.1111/jtsa.12707

Seo, W.-K. (2023b). Functional Principal Component Analysis for Cointegrated Functional Time Series.

Weidmann, J. (2012). *Linear operators in Hilbert spaces* (Vol. 68). Springer Science & Business Media.

Wold, H. (1966). Estimation of principal components and related models by iterative least squares. *Multivariate analysis*, 391-420.

Wold, H. (1973). Nonlinear iterative partial least squares (NIPALS) modelling: some current developments. En *Multivariate analysis–III* (pp. 383-407). Elsevier.

Wold, H. (1979). Model construction and evaluation when theoretical knowledge is scarce: An example of the use of partial least squares cahiers du departement deconometrie. *Geneva, Switzerland: Faculté des Sciences Économiques et Sociales, Université de Genève*.

Yosida, K. (2012). *Functional analysis* (Vol. 123). Springer Science & Business Media.