

Identificación de Sistemas Dinámicos : Un Enfoque Integrador Basado En Esparcimiento, Regularización y Aproximación de Bajo Rango



Kerin Omar Cardona Mejía

Universidad Nacional Autónoma de Honduras
Facultad de Ciencias
Maestría en Ingeniería Matemática
Tegucigalpa M.D.C., Honduras
23 de octubre de 2023

Identificación de Sistemas Dinámicos : Un Enfoque Integrador Basado En Esparsamiento, Regularización y Aproximación de Bajo Rango

Kerin Omar Cardona Mejía

Trabajo de grado presentado como requisito parcial para optar por el título de:

Master en Matemática con Orientación en Ingeniería Matemática

Director:

Dr. Fredy Antonio Vides

Grupo de investigación:

Centro de Innovación y Cálculo Científico (CICC-UNAH)

Universidad Nacional Autónoma de Honduras

Facultad de Ciencias

Maestría en Matemática con Orientación en Ingeniería Matemática

Tegucigalpa M.D.C., Honduras

23 de octubre de 2023

Dedicatoria

*A mi esposa Scarleth y mis padres Teodoro y Elisa , por su apoyo incondicional en
este largo camino.*

Agradecimientos

En primer lugar agradecer al Creador del Universo, Dios, por darme la sabiduría y fortaleza para poder culminar este trabajo. Reconozco que de Él emana todo conocimiento y voluntad.

A Fredy Vides, por haber compartido sus ideas, consejos y tiempo en todo el proceso de desarrollo de este trabajo de tesis. Agradecerle por demostrar siempre una gran ética de trabajo y disponibilidad para compartir de forma abierta sus conocimientos y experiencias en el duro proceso de investigación científica. Además por permitirme dar mis primeros pasos como investigador en el Centro de Innovación en Computación Científica (CICC-UNAH).

A mi esposa Scarleth, por sacrificar nuestro valioso tiempo juntos y por servir de fortaleza en los momentos más complicados en este proceso. De igual forma a mis padres, Teodoro y Elisa, por darle la oportunidad de ingresar a la universidad y sacar adelante mi carrera.

Por último, pero no menos importante, a las autoridades y catedráticos de la Maestría en Matemáticas UNAH, por el gran esfuerzo de impulsar el proceso de desarrollo de la matemática en Honduras.

Resumen

En este proyecto se presentan aplicaciones de algunas técnicas teóricas y computacionales, para la aproximación regularizada y esparcida de sistemas dinámicos basados en datos. La investigación abordada en este proyecto abarca modelos lineales, con aplicaciones en ingeniería, ciencia y econometría. Específicamente, las propiedades de regularización y esparcimiento se integran en la aproximación de los parámetros del modelo, especialmente en el contexto de la aproximación de bajo rango. Es importante señalar que los resultados obtenidos son independientes de una representación específica del sistema y no asumen particiones de datos de entrada y salida. Estas metodologías antes mencionadas son sujetas a comparación mediante simulaciones numéricas y aplicaciones de conjuntos de datos del mundo real.

Palabras clave: Identificación de sistemas dinámicos, esparcimiento y regularización, aproximación por bajo rango, series de tiempo.

Abstract

This project presents applications of some theoretical and computational techniques for the regularized and sparse approximation of data-driven dynamic systems. The research undertaken in this project encompasses linear models, with applications in engineering, science, and econometrics. Specifically, the properties of regularization and sparsity are integrated into the parameter approximation of the model, especially in the context of low-rank approximation. It is important to note that the obtained results are independent of a specific system representation and do not assume partitions of input and output data. These aforementioned methodologies are subject to comparison through numerical simulations and real-world dataset applications.

Palabras clave: System identification, regularization and sparsity, low rank approximation, time series analysis.

Contenido

Contenido	III
Lista de figuras	VI
Lista de tablas	VII
1. Introducción	1
2. Preliminares y Notación	4
2.1. Sistemas Dinámicos	4
2.2. Sistemas Lineales Invariantes en el Tiempo	9
2.3. Identificación de Sistemas Dinámicos	10
2.4. Esparcimiento y Regularización	11
2.5. Validación y Criterio de Selección	12
3. Métodos	14
3.1. Aproximación de Bajo Rango	14
3.2. Mínimos Cuadrados Esparcido con SDSI	15
3.3. Identificación Esparcida con SpLra	17
3.4. Modelos Para Series de Tiempo	19
3.5. Reducción de Dimensión y Modelos Factoriales Dinámicos	24
4. Experimentos y Resultados	26
4.1. Identificación de un Sistema SISO	28

<i>CONTENIDO</i>	III
4.2. Identificación de un Sistema MIMO	31
4.3. Aplicaciones En Procesos Industriales	35
4.4. Nivel de Agua Lago Erie	40
4.5. Aplicaciones En Econometría	43
5. Discusión y Trabajo Futuro	51
5.1. Conclusiones	51
5.2. Trabajo Futuro	52
Bibliografía	57

Lista de figuras

2.1. Bloque básico de un sistema dinámico.	4
4.1. Serie de tiempo para la fase de identificación.	29
4.2. Polos y ceros del sistema simulado original.	29
4.3. Serie de tiempo para la fase de validación.	30
4.4. Serie de tiempo del sistema perturbado con una estructura ARMAX con $\sigma^2 = 0.010$	31
4.5. Señales de entrada PBRS.	32
4.6. Señales de salida generadas a partir del sistema original.	33
4.7. Señales de salida del sistema original a partir de un segundo grupo de señales de entrada. Esto datos simulados se utilizan en el proceso de validación.	33
4.8. Salida gráfica del desajuste, para diferentes valores del orden del mo- delo. La grafica nos sugiere utilizar un valor de $L = 5$	34
4.9. Señales de salida del sistema original e identificada.	34
4.10. Señales de salida original e identificada en la fase de validación.	34
4.11. Salida gráfica de los ceros del sistema identificado.	35
4.12. Salida gráfica de los datos de entrada y salida de un brazo robótico flexible	37
4.13. Salida gráfica de los últimos 200 datos de salida originales e identifi- cados	37
4.14. Salida gráfica de los datos de entrada y salida de un secador	38

4.15. Salida gráfica de los últimos 300 datos de salida originales e identificados	38
4.16. Salida gráfica de los primeros 500 datos(s) de salida originales	39
4.17. Salida gráfica de los últimos 500 datos(s) de salida originales e identificados	39
4.18. Niveles de agua Lago Erie, 1921-1970.	40
4.19. Salida gráfica de los datos sin tendencia y estandarizados. La parte sombreada representa el 30% del total para el proceso de validación.	41
4.20. Comportamiento del <i>rmse</i> de los datos de identificación y validación con respecto al orden ($L=lag$) del modelo.	41
4.21. (a) Residuos del modelo autor-regresivo AR(3) con respecto a la serie original de datos. (b) Histograma de los residuos ϵ_i estandarizados. (c) Q-Q plot.	42
4.22. Niveles de agua Lago Erie, 1921-1970.	43
4.23. Serie de tiempo tasa de desempleo <i>UNRATE</i> en USA, 2008-2022.	44
4.24. Salida gráfica de nueve variables macroeconómicas seleccionadas originalmente.	45
4.25. Salida gráfica de nueve variables macroeconómicas con tendencia lineal removida y estandarizadas.	45
4.26. Desajuste (<i>misfit</i>) del modelo para diferentes valores de L . La gráfica nos indica que aumentar el orden del modelo no mejora la diferencia entre los errores en los conjuntos de identificación y validación.	46
4.27. Salida gráfica del modelo identificado con <code>SpLra</code>	47
4.28. Salida gráfica del modelo identificado con <code>n4sid</code>	47
4.29. Salida gráfica del modelo identificado con <code>AutoReg</code>	48
4.30. Salida gráfica de la serie <i>UNRATE</i> original (azul) y la serie generada a partir de reducción de orden con $r = 9$ factores latentes (rojo). La parte sombreada corresponde a los datos seleccionados para la fase de validación.	49

4.31. Salida gráfica de los factores $F \in \mathbb{R}^{9 \times 756}$ calculados a partir del Algoritmo 4.	49
4.32. El gráfico muestra el <i>rmse</i> en el conjunto de identificación (azul) y validación (rojo), para diferentes valores de L . Siguiendo con el método heurístico de la selección del orden del modelo, se toma $L = 6$.	50
4.33. Salida gráfica de la serie UNRATE original (azul) y la serie generada a partir de reducción de orden con $r = 9$ factores latentes (rojo). La parte sombreada corresponde a los datos seleccionados para la fase de validación.	50

Lista de tablas

4.1. Resumen de resultados.	30
4.2. Función de transferencia sistema MIMO original.	32
4.3. Resumen de resultados sistema MIMO.	33
4.4. Selección de conjuntos de datos DAISY.	36
4.5. Resultados base de datos Daisy.	40
4.6. Selección de variables macroeconómicas.	44
4.7. Resumen de resultados modelo econométrico base de datos FRED- MD.	48

Capítulo 1

Introducción

La construcción de modelos basados en datos es una pieza fundamental en aplicaciones en ciencias e ingeniería. La identificación de sistemas consiste en utilizar técnicas desarrolladas para construir modelos matemáticos de sistemas dinámicos a partir de datos observados.

Es de gran interés en la actualidad, la identificación de modelos de alta dimensión, los cuales han ganado una importante relevancia en muchas áreas, por ejemplo, la economía [8], finanzas [10] y aprendizaje automático [18]. Por ejemplo, los modelos autorregresivos vectoriales VAR (por sus siglas en inglés), son considerados la clave para el análisis de información estructurada en la evolución conjunta de series de tiempo macroeconómicas.

Los modelos VAR estándar no suelen incluir más de diez variables [4], sin embargo es de interés para científicos, ingenieros y econométristas poder modelar cientos de series de datos. Dado que el número de parámetros crece de manera cuadrática con respecto al tamaño del modelo, es necesario explorar alternativas a los modelos VAR, por ejemplo, los modelos de factores aumentados FAVAR [19]. Bernanke utilizó modelos FAVAR en [5], para estimar los efectos de las políticas monetarias.

Algunos ejemplos de aplicación de modelos multivariados de alta dimensión, son los datos de tipo panel a gran escala [20], matrices de volatilidad [14], riesgo y

construcción de portafolio y modelos de riesgo crediticio. En general los problemas de alta dimensión proponen un gran reto en la estimación de parámetros, debido a que pequeños errores en la estimación de los elementos dan como resultado grandes errores en las matrices.

Los modelos esparcidos representan una de las mejores alternativas para tratar problemas de alta dimensión [18]. La idea principal es asumir que el vector de parámetros contiene muchas componentes exactamente cero o lo suficientemente pequeñas. Dicha suposición es de gran importancia para lograr la identificación del verdadero modelo, especialmente si solo se cuenta con una muestra de datos pequeña.

Extraer información de datos económicos, tomando en cuenta características especiales de dichos datos y la información a priori de la teoría económica, es uno de los objetivos principales que persigue la econometría [17, 16]. El estudio de series temporales y ecuaciones en diferencias se han convertido en una fuente de inspiración para el desarrollo de la estadística en los últimos años.

Debido al gran avance en las ciencias computacionales, los datos de alta dimensión han surgido en muchos campos de la ciencia, ingeniería, humanidades y la economía. Diversos temas como ser comercio electrónico, mercadeo, genética, imágenes biomédicas y muchos más, proporcionan datos a gran escala que requieren de nuevas técnicas teóricas y computacionales para un tratamiento más efectivo.

Los sistemas dinámicos aplicados a los mercados financieros actualmente reciben mucha atención a medida que la información de los mercados se vuelve disponible [14]. Construir modelos de los mercados financieros es un problema inverso, dado que no se cuenta con principios fundamentales, como en el caso de la física e ingeniería.

El propósito de la investigación reportada en este documento, es aplicar algunas técnicas teóricas y computacionales presentadas en [32, 24, 33, 35], para la aproximación de sistemas basados en datos de alta dimensión. Nuestro enfoque consiste en integrar los métodos y algoritmos en dichos trabajos, que nos permitan identificar,

basados en datos, los modelos que posteriormente pueden ser usados para un análisis detallado de la estructura y dinámica del fenómeno.

En el Capítulo 2 se presenta la notación básica y los conceptos preliminares. Se establecen las definiciones de sistema dinámico en tiempo discreto, el marco teórico general de identificación de sistemas, así como la estructura y representación de los modelos considerados en este trabajo.

En el Capítulo 3 aborda los métodos para resolver el problema principal de este estudio: identificación esparcida y regularizada de modelos lineales por medio de la aproximación de bajo rango. Además se introducen los algoritmos principales utilizados en esta investigación: *SDSI* (Sparse Dynamic System Identification) y *SpLra* (Sparse low rank approximation).

En el Capítulo 4, se valida la solidez tanto teóricas como computacional mediante el uso de datos simulados, así como a través de la base de datos FRED-MD [25] y la base de datos Daisy [13]. Esta es una parte fundamental del presente trabajo, poder mostrar como la aproximación de bajo rango ayudan a tratar el problema de reducción de dimensión. En la parte final se presentan las conclusiones y trabajo futuro.

Capítulo 2

Preliminares y Notación

2.1. Sistemas Dinámicos

En términos generales un *sistema dinámico* es un objeto donde diferentes clases de variables se combinan para producir una *salida*. Estas señales de salida pueden ser afectadas por otras señales externas que pueden ser manipuladas por el observador llamadas *entradas*. También pueden ser señales no observables que distorsionan tanto las señales de entrada como de salida.

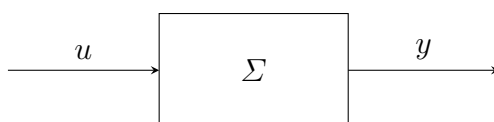


Figura 2.1: Bloque básico de un sistema dinámico.

El enfoque anterior de un sistema dinámico es ampliamente utilizado por la comunidad científica, sin embargo conduce hacia algunos inconvenientes detallados ampliamente en [37, 38, 36]. Uno de las desventajas es el hecho de asumir una partición de las variables en tipo de entrada y salida, lo cual puede conducir a resultados imprecisos. En el mismo trabajo mencionado anteriormente se presenta un enfoque alternativo conocido como *enfoque del comportamiento*, donde se corrige y define de manera mas general y precisa un sistema dinámico.

De manera más precisa, un modelo dinámico con q variables es un subconjunto del espacio de trayectorias $(\mathbb{R}^q)^{\mathbb{T}}$, el conjunto de todas las funciones desde \mathbb{T} al espacio de variables \mathbb{R}^q [24].

Definición 1. *Un sistema dinámico Σ se define como una tripleta $\Sigma = (\mathbb{T}, \mathbb{W}, \mathcal{B})$, donde $\mathbb{T} \subset \mathbb{R}$ es el eje del tiempo, \mathbb{W} es el espacio de las señales y $\mathcal{B} \subset \mathbb{W}^{\mathbb{T}}$ es el comportamiento.*

La forma usual de representar el comportamiento $\mathcal{B} \subset \mathbb{W}^{\mathbb{T}}$ de un sistema dinámico es mediante ecuaciones, por ejemplo, el conjunto de funciones de la forma $f : \mathbb{W}^{\mathbb{T}} \rightarrow \mathbb{R}^q$ o $\mathcal{B} = \{w \in \mathbb{W}^{\mathbb{T}} \mid f(w) = 0\}$

En este estudio nos centraremos en el caso $\mathbb{T} = \mathbb{Z}$, que se conoce como *sistema dinámico discreto*. Cada observación $w \in \mathbb{W}^{\mathbb{T}}$ es llamada tradicionalmente como *serie de tiempo*.

Definición 2. *Un sistema dinámico discreto Σ , con variables de entrada y salida $u \in \mathbb{R}^m$, $y \in \mathbb{R}^n$ respectivamente, se define como la tripleta $(\mathbb{Z}, \mathcal{X}, \mathcal{F})$, determinado por un conjunto de estados $\mathcal{X} \in \mathbb{R}^N$ y una función $\mathcal{F} : \mathcal{X} \rightarrow \mathbb{R}^n$ llamado operador de transición.*

De tal forma que para cualquier *serie de tiempo* determinada por la secuencia $\{w_t = (u_t, y_t)\}_{t \geq 1} \subset \mathbb{R}^{m+n}$, se tiene que $\mathcal{F}(y_t, u_t) = y_{t+1}$.

Ejemplo 1. *Un sistema dinámico lineal autónomo (sin variables de entrada) es un modelo simple para la serie $\{x_t\} \subset \mathbb{R}^n$, en el cual cada x_{t+1} está dado por*

$$x_{t+1} = Ax_t, \quad t = 1, 2, \dots \quad (2.1)$$

donde $A \in \mathbb{R}^{n \times n}$ es llamada la matriz dinámica del sistema.

El modelo (2.1) se puede extender a esquemas más generales agregando variables exógenas, es decir, variables que provienen desde fuera del sistema, por ejemplo

$$x_{t+1} = Ax_t + Bu_t + c_t, \quad t = 1, 2, \dots \quad (2.2)$$

donde $u_t \in \mathbb{R}^m$ es un vector de entradas, $B \in \mathbb{R}^{n \times m}$ es una matriz de entrada y $c_t \in \mathbb{R}^n$ es un vector de compensación.

Ejemplo 2 (Modelo de Markov [7]). *Un modelo lineal de Markov de orden K , donde el estado x_{t+1} depende del estado actual y de los $K - 1$ estados anteriores, se puede expresar de la forma*

$$x_{t+1} = A_1 x_t + \dots + A_K x_{t-K+1}, \quad t = K, K + 1, \dots \quad (2.3)$$

Ejemplo 3 (Función de Transferencia [23]). *El sistema \mathfrak{S} , con parámetros la matriz racional $H \in \mathbb{R}^{n \times m}(z)$ se puede representar de la forma*

$$\mathfrak{S}_{i/o}(H) := \left\{ \begin{bmatrix} u \\ y \end{bmatrix} \in (\mathbb{R}^{m+n})^{\mathbb{N}} \mid \mathcal{F}(y) = H(z)\mathcal{F}(u) \right\} \quad (2.4)$$

donde \mathcal{F} es la transformada Z [1].

Los modelos de la forma (2.3) son muy utilizados en el análisis de series de tiempo y modelos econométricos. En la sección 3.4 se presentan brevemente los tipos de modelos clásicos auto-regresivos y su conexión con los sistemas dinámicos lineales.

Las variables de estados son un tipo especial de variables que especifican la memoria dinámica de un sistema. En general, un modelo de espacio de estados presenta dos características esenciales. En primer lugar existe un proceso oculto o latente x_t que se conoce como variable de estados. La segunda característica es que dicho proceso es de Markov o auto-regresivo.

Ejemplo 4 (Modelo espacio de estados). *El sistema \mathfrak{S} , inducido por ecuaciones de entrada (u)/estado (x)/salida (y) con parámetros A, B, C, D se puede representar de la forma*

$$\mathfrak{S}_{i/s/o}(A, B, C, D) := \left\{ \begin{bmatrix} u \\ y \end{bmatrix} \in (\mathbb{R}^{m+n})^{\mathbb{N}} \mid \exists x \in (\mathbb{R}^n)^{\mathbb{N}} \right\}$$

tal que

$$\begin{aligned}x_{t+1} &= Ax_t + Bu_t, \\y_t &= Cx_t + Du_t.\end{aligned}$$

Un caso particular de los modelos de espacio de estados son los modelos *factoriales dinámicos*, los cuales son muy utilizados para hacer pronósticos en macroeconomía. Estos modelos se construyen de forma que solo dependen de un número esencial de factores no observables. Los modelos factoriales descomponen el comportamiento de una variable económica x_{it} en componentes controladas por algunos factores f_t , los cuales son comunes para todas las variables, pero con efectos específicos en cada una de ellas [31]. La variable a pronosticar depende de una combinación lineal de sus valores pasados y también depende de los siguientes factores a estimar

$$x_t = \bar{x}_t + Af_t, \quad (2.5)$$

donde x_t es el vector de observaciones en el tiempo t , \bar{x}_t es el promedio de x_t , $A = (\lambda_1, \dots, \lambda_N)^T$ es una matriz de cargas de $N \times r$ y f_t es el vector de factores de $r \times 1$. Generalmente en modelos macroeconómicos la dimensión de x_t es mayor que la dimensión de f_t . Además, los factores son modelados como un proceso VAR(L) de la forma

$$\hat{f}_{t+1} = A_1 f_t + \dots + A_L f_{t-L}, \quad (2.6)$$

donde A_1, \dots, A_L son matrices de coeficientes auto-regresivos.

Definición 3. Dada una serie de tiempo $\Sigma = \{x_t\} \subset \mathbb{R}^n$, un entero positivo L ,

denotaremos la matriz de trayectorias tipo Hankel en bloques correspondiente a Σ_T

$$\mathcal{H}_L(\Sigma_T) = \begin{bmatrix} x_1 & x_2 & x_3 & \cdots & x_{T-L+1} \\ x_2 & x_3 & x_4 & \cdots & x_{T-L+2} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ x_L & x_{L+1} & x_{L+2} & \cdots & x_T \end{bmatrix}$$

Teorema 1 (Descomposición en Valores Singulares [15]). *Sea $A \in \mathbb{R}^{m \times n}$, existen matrices ortogonales $U \in \mathbb{R}^{m \times m}$ y $V \in \mathbb{R}^{n \times n}$ tales que*

$$U^T A V = \Sigma = \text{diag}(\sigma_1, \dots, \sigma_p) \in \mathbb{R}^{m \times n},$$

donde $p = \min\{m, n\}$, además $\sigma_1 \geq \sigma_2 \dots \geq \sigma_p$.

En el resto de este documento, nos referiremos a la descomposición anterior como SVD (por sus siglas en inglés). Sea $A = U \Sigma V^T \in \mathbb{R}^{m \times n}$ la SVD de A con $m \geq n$, entonces

$$A = U_1 \Sigma_1 V^T$$

donde $U_1 = U(:, 1 : n) \in \mathbb{R}^{m \times n}$ y $\Sigma_1 = \text{diag}(\sigma_1, \dots, \sigma_n) \in \mathbb{R}^{n \times n}$. Nos referiremos a la versión truncada de la SVD como *SVD reducida*.

Definición 4. *Dado $\delta > 0$ y una matriz $X \in \mathbb{R}^{m \times n}$, denotaremos el rango de X por $\text{rk}_\delta(X) > 0$ al valor determinado por*

$$\text{rk}_\delta(X) = \sum_{i=1}^p \sigma_i, \quad \sigma_i > \delta,$$

donde los valores σ_i corresponden a los valores singulares de la SVD reducida de A y $p = \min\{m, n\}$.

La presencia de alta dimensión conlleva un gran reto en el desarrollo de la teoría econométrica. Debido a la habilidad de resumir información de conjuntos de datos grandes, los modelos factoriales se han convertido en una herramienta indispensable

en economía y finanzas. Una revisión importante del aporte de los modelos factoriales en estimación e inferencia estadística se puede encontrar en [3].

Bernanke, Boivin y Eliasch investigaron en [5] soluciones a problemas que surgieron en la implementación de los modelos VAR. Estos problemas se deben a la información limitada que generalmente se utiliza en estos modelos empíricos. La metodología implementada FAVAR, es decir, VAR con factor aumentado, fue aplicada con éxito en la identificación de los efectos de política monetaria en la economía.

Durante la última década los métodos basados en datos se han desarrollado de gran manera, esto debido a la gran cantidad de datos y evolución de técnicas analíticas. En [28] se estudian modelos cuadráticos esparcidos basados en polinomios, en los cuales la dinámica depende un número reducido de variables e interacciones de segundo orden. En particular en [28] se detallan estrategias de muestreo que conllevan a la identificación exacta de sistemas dinámicos de primer orden, aun cuando solo se cuenta con un número reducido de muestras.

2.2. Sistemas Lineales Invariantes en el Tiempo

Los sistemas *lineales invariantes en el tiempo* LTI (por sus siglas en inglés), forman la clase de sistemas dinámicos más utilizados en la práctica. Lo anterior se debe por los excelentes resultados en la mayoría de los casos aplicados a la vida real.

- Un modelo \mathcal{B} es *lineal* si es un subespacio del espacio de trayectorias $(\mathbb{R}^q)^{\mathcal{T}}$.
- Un modelo \mathcal{B} es *invariante en el tiempo* si $\sigma^\tau \mathcal{B} = \mathcal{B}$, donde

$$(\sigma^\tau w)(t) := w(t + \tau), \quad \text{para todo } t \in \mathcal{T}.$$

- Si el modelo \mathcal{B} es de dimensión finita, el comportamiento futuro del modelo es determinísticamente relacionado con un vector de *estado* de dimensión finita.
- Denotaremos por \mathcal{L}^w la clase de todos los sistemas LTI con w variables, es

decir el espacio de señales $\mathbb{W} = \mathbb{R}^v$.

En este proyecto de investigación sobre identificación de sistemas dinámicos y aplicaciones a series temporales, los sistemas *LTI* nos permiten modelar y predecir de manera efectiva la evolución de datos temporales, lo que puede ser crucial en áreas como la predicción económica, el pronóstico del clima o el control de procesos industriales.

2.3. Identificación de Sistemas Dinámicos

Existen muchas maneras de intentar determinar la dinámica de un sistema, una idea generalizada es construir modelos basados en principios físicos del sistema. Este enfoque puede no ser factible en el caso donde dichos principios fundamentales son desconocidos, por ejemplo, el comportamiento humano, economía, consumo de energía, etc.

Un enfoque diferente es utilizar datos o mediciones de las variables del sistema para construir un modelo del sistema, este proceso se conoce como *identificación de sistemas*. En un mundo cada vez más orientado hacia los datos, este tipo de modelos están obteniendo mayor interés en los últimos años. Otra ventajas, además de las ya mencionadas, es el hecho de no requerir conocimiento previo o entendimiento del proceso.

El proceso de identificación de sistemas lo podemos resumir en tres componentes principales [30]: el conjunto de datos obtenido a partir de mediciones de entrada-salida, el conjunto de modelos candidatos o estructura del modelo y el criterio de selección para el mejor modelo del conjunto de candidatos y una regla para evaluar los modelos candidatos basado en datos.

Problema 1 (Identificación de Sistemas en General). *Dada la serie de tiempo $w \in \mathbb{R}^{(m+n) \times N}$, posiblemente representando mediciones de entrada-salida de la forma $w = \begin{bmatrix} u & y \end{bmatrix}^T$, de un sistema desconocido \mathfrak{S} , encontrar un modelo razonable para*

el sistema basado en los datos de las mediciones u, y .

Definición 5 (Representación del núcleo de un sistema dinámico [33, 23]). *Dado un conjunto de trayectorias \mathcal{D} correspondientes a resultados de un experimento \mathbb{E} , en un conjunto universal \mathcal{U} de posibles resultados. Denominaremos modelo dinámico lineal correspondiente al experimento \mathbb{E} , al conjunto $\mathcal{M}_{\mathbb{E}}$ representado por la siguiente expresión.*

$$\mathcal{M}_{\mathbb{E}} := \{d \in \mathcal{U} : f_E(d) = 0\}, \quad (2.7)$$

donde $f_E \in (\mathbb{R}^n)^{\mathcal{U}}$ es una función de \mathcal{U} a \mathbb{R}^n .

2.4. Esparcimiento y Regularización

Con la disponibilidad de datos de alta dimensión, es indispensable contar con herramientas que permitan seleccionar las variables de mayor relevancia. En ese sentido, los métodos de regularización se ha convertido en una poderosa herramienta para la estimación y selección de modelos.

Dados $X \in \mathbb{R}^{m \times n}$ y $y \in \mathbb{R}^m$, el problema de aproximación de mínimos cuadrados consiste en

$$\beta = \arg \min_{\hat{\beta} \in \mathbb{R}^n} \|y - X\hat{\beta}\|_2, \quad (2.8)$$

Una solución conocida es utilizar la pseudo-inversa de X , para lo cual el vector solución de 2.8 es

$$\beta = X^\dagger b = (X^T X)^{-1} X^T y \quad (2.9)$$

En el caso que X esté mal condicionada, el cálculo de X^\dagger se vuelve inestable y esto puede resultar en problemas de sobre ajuste en el modelo [6]. Para contrarrestar las situaciones de inestabilidad numérica en la práctica se utilizan técnicas de regularización.

El proceso de regularización conocido como Lasso, es una extensión del método convencional de mínimos cuadrados (OLS, por sus siglas en inglés). Con esta técnica

se espera una mejor aproximación de los coeficientes de forma esparcida.

$$\beta = \arg \min_{\hat{\beta} \in \mathbb{R}^n} \left\{ \|y - X\hat{\beta}\|_2 + \lambda \|\beta\|_1 \right\}, \quad (2.10)$$

donde $\lambda > 0$ se conoce como parámetro de regularización.

Otra técnica de regularización importante consiste en utilizar la descomposición truncada en valores singulares SVD, luego el problema resulta en

$$\beta = \min_{\hat{\beta} \in \mathbb{R}^n} \|y - X_\delta \hat{\beta}\|_2, \quad (2.11)$$

donde $\delta < \text{rank}(X)$ es un parámetro y $X_\delta \in \mathbb{R}^{m \times n}$ es la mejor aproximación de rango δ de X calculada a partir de la SVD.

En este documento, utilizaremos resultados y algoritmos presentados en [33], que nos permiten encontrar una solución al problema 2.11. La idea principal es utilizar métodos de aproximación de bajo rango y reducción de orden para la solución del problema lineal de mínimos cuadrados esparcido, esto es, soluciones con un número máximo de entradas diferentes de cero.

2.5. Validación y Criterio de Selección

Una de las características más importantes de un modelo es la habilidad para generalizar, es decir, no solo lograr buen desempeño en un conjunto de datos dado, sino también obtener un desempeño razonable en datos nuevos. Un criterio para verificar una buena capacidad de generalizar el ajuste de un modelo es comparar predicciones de nuevos datos, con las predicciones logradas en los datos utilizados para construir el modelo.

En general se desea que el modelo tenga desempeños similares en ambos casos, si ocurre lo contrario diremos que el modelo tiene poca habilidad de generalización. Un método simple de validación es dividir los datos en dos subconjuntos: un conjunto

de entrenamiento (entre el 50 %-90 %) y el resto, conjunto de prueba completamente oculto para validar el modelo en datos nuevos.

En resumen, se quiere encontrar modelos que hagan buenas predicciones en el conjunto de entrenamiento y también buenas predicciones en el conjunto de prueba. Con el objetivo de identificar un buen modelo candidato, se utiliza la identificación por mínimos cuadrados para varios modelos con diferentes valores de L . Luego se debe escoger el modelo que tenga el menor error en la predicción y también sea consistente en los datos de entrenamiento y prueba.

Capítulo 3

Métodos

En esta sección, se presentan definiciones, resultados y algoritmos que nos permiten integrar las propiedades de regularización y esparcimiento en la aproximación de los parámetros de un sistema. Como punto de partida, se construye a partir de los datos la matriz de trayectorias tipo `hankel`. El rango de esta matriz se convierte en el factor determinante que la conecta con el problema de identificación de sistemas, es decir, para pertenecer a la clase de sistemas LTI. Adicionalmente utilizamos `SDSI` y `SpLra` para forzar la regularización y el esparcimiento en los parámetros.

3.1. Aproximación de Bajo Rango

El objetivo de la aproximación de una matriz por otra matriz con rango menor o igual al original, es obtener una representación reducida de los datos con una mínima pérdida de información. Una colección de datos es almacenada en $D \in \mathbb{R}^{n \times N}$, donde cada columna de D corresponde a un vector en \mathbb{R}^n .

El problema de modelación aproximada consiste en que dado un conjunto de puntos en \mathbb{R}^n , encontrar un subespacio de \mathbb{R}^n de dimensión acotada que minimice la distancia al conjunto de datos.

Problema 2 (Aproximación de bajo rango). *Dada $X \in \mathbb{R}^{n \times N}$, donde $n \leq N$, un*

entero r , $0 < r < n$, encontrar

$$\widehat{X}^* := \arg \min_{\widehat{X}} \|X - \widehat{X}\|_F \quad \text{sujeto a: } \text{rank}(\widehat{X}) \leq r.$$

Teorema 2 (Eckart-Young-Mirsky). Sea $X = U\Sigma V^T$ la descomposición reducida en valores singulares de X y la siguiente partición en bloques

$$U := \begin{bmatrix} r & q-r \\ U_1 & U_2 \end{bmatrix}_q, \quad \Sigma := \begin{bmatrix} r & q-r \\ \Sigma_1 & 0 \\ 0 & \Sigma_2 \end{bmatrix}_{q-r}, \quad V := \begin{bmatrix} r & q-r \\ V_1 & V_2 \end{bmatrix}_N,$$

entonces una solución al Problema 2 es $\widehat{X}^* = U_1 \Sigma_1 V_1^T$.

Además el error de aproximación es $\|X - \widehat{X}^*\|_F = \sqrt{\sigma_{r+1}^2 + \dots + \sigma_n^2}$.

Lo anterior implica que es posible aproximar una matriz de datos de alta dimensión por r términos dominantes de las columnas de U_1 y V_1 de la forma

$$\widehat{X}^* = \sum_{k=1}^r \sigma_k u_{1,k} v_{1,k}^T = \sigma_1 u_{1,1} v_{1,1}^T + \sigma_2 u_{1,2} v_{1,2}^T + \dots + \sigma_r u_{1,r} v_{1,r}^T. \quad (3.1)$$

Algoritmo 1: Aproximación de bajo rango lra

Datos: $X \in \mathbb{R}^{m \times n}$, $r \in \mathbb{Z}^+$

Resultado: $X_r = \text{lra}(X, r)$

$U, S, V \leftarrow \text{svd}(X)$, descomposición reducida

$s \leftarrow \min\{m, n\}$

$U_r \leftarrow \sum_{j=1}^r U \hat{e}_{j,s} \hat{e}_{j,s}^T$, $S_r \leftarrow \sum_{j=1}^r \hat{e}_{j,s}^T S \hat{e}_{j,s}$, $V_r \leftarrow \sum_{j=1}^r \hat{e}_{j,s} \hat{e}_{j,s}^T V$

$X_r \leftarrow U_r S_r V_r^T$

3.2. Mínimos Cuadrados Esparcido con SDSI

El siguiente resultado nos permiten encontrar soluciones regularizadas y esparcidas a tipos de problemas de la forma 2.8. En general, nos interesa el problema extendido cuando el vector y en 2.8 se convierte en una matriz. Es decir el problema

matricial de mínimos cuadrados.

Problema 3 ([33]). *Dados $\delta > 0$ y dos matrices $X \in \mathbb{R}^{m \times n}$, $Y \in \mathbb{R}^{m \times p}$, denotaremos por $Y \approx_\delta XA$ para representar el problema de encontrar $A \in \mathbb{R}^{n \times p}$, $\alpha, \beta \geq 0$ y un proyector ortogonal Q tal que*

$$\|Y - XA\|_F \leq \alpha\delta + \beta\|(I_m - Q)Y\|_F$$

La matriz A la llamaremos una solución del problema $Y \approx_\delta XA$.

Teorema 3 ([33]). *Dado $\delta > 0$ y dos matrices $X \in \mathbb{R}^{m \times n}$, $Y \in \mathbb{R}^{m \times p}$. Si $r = \text{rk}(A)_\delta > 0$ entonces, existe una solución A al problema $Y \approx_\delta XA$ con a lo sumo rp entradas diferentes de cero.*

Las soluciones descritas en el resultado anterior, se pueden obtener utilizando el siguiente algoritmo. Una implementación computacional del Algoritmo 2 la podemos encontrar en [34, 12] como `lsspsolver`.

Algoritmo 2: Mínimos Cuadrados Lineal Esparcido [33]

Datos: $X \in \mathbb{R}^{m \times n}$, $Y \in \mathbb{R}^{m \times p}$, $\delta > 0$, $N \in \mathbb{Z}^+$, $\epsilon > 0$
Resultado: $A = \text{lspsolver}(X, Y, \delta, N, \epsilon)$
 $U, S, V \leftarrow \text{svd}(X)$, descomposición reducida

 $s \leftarrow \min\{m, n\}$, $r \leftarrow \text{rk}_\delta(X)$
 $U_\delta \leftarrow \sum_{j=1}^r U \hat{e}_{j,s} \hat{e}_{j,s}^T$, $T_\delta \leftarrow \sum_{j=1}^r (\hat{e}_{j,s}^T S \hat{e}_{j,s})^{-1} \hat{e}_{j,s} \hat{e}_{j,s}^T$, $V_\delta \leftarrow \sum_{j=1}^r \hat{e}_{j,s} \hat{e}_{j,s}^T V$
 $\hat{X} \leftarrow U_\delta^T X$, $\hat{Y} \leftarrow U_\delta^T Y$
 $A_0 \leftarrow V_\delta^T T_\delta \hat{Y}$
para $j = 1, \dots, p$ **hacer**

 | $K \leftarrow 1$, error $\leftarrow 1 + \delta$

 | $c \leftarrow A_0 \hat{e}_{j,p}$, $x_0 = c$

 | $\hat{c} \leftarrow \begin{bmatrix} \hat{c}_1 & \dots & \hat{c}_n \end{bmatrix}^T = \begin{bmatrix} |\hat{e}_{1,n}^T c| & \dots & |\hat{e}_{n,n}^T c| \end{bmatrix}^T$

 | Calcular la permutación $\sigma : \{1, \dots, n\} \rightarrow \sigma : \{1, \dots, n\}$ tal que:

 | $\hat{c}_{\sigma(1)} \geq \hat{c}_{\sigma(2)} \geq \dots \geq \hat{c}_{\sigma(n)}$

 | $N_0 \leftarrow \max \left\{ \sum_{j=1}^n H_\epsilon(\hat{c}_{\sigma(j)}), 1 \right\}$

 | **mientras** $K \leq N$ **and** error $> \delta$ **hacer**

 | | $x \leftarrow \mathbf{0}_{n,1}$

 | **fin**
fin

3.3. Identificación Esparcida con SpLra

Definición 6 (Representación del Núcleo de un Sistema Dinámico Lineal [33, 23]).

Dada una serie de tiempo $w \in \mathbb{R}^{q\mathbb{Z}}$ correspondientes a resultados de un experimento \mathbb{E} , denominaremos modelo dinámico lineal correspondiente al experimento \mathbb{E} , al conjunto $\mathcal{B}_{\mathbb{E}}$ representado por la siguiente expresión

$$\mathcal{B}_{\mathbb{E}} := \{w \in \mathbb{R}^{q\mathbb{Z}} : R(\sigma) = 0\}, \quad (3.2)$$

donde $R(\sigma) = R_0 + R_1\sigma + \dots + R_L\sigma^L$ es una matriz polinomial de grado L y las matrices R_0, R_1, \dots, R_L toman valores en $\mathbb{R}^{q-m \times n}$.

Problema 4 (Identificación exacta). *Dada la serie de tiempo $w \in \mathbb{R}^{m+n}$, generada a partir de un sistema desconocido \mathfrak{S} y el orden del sistema $\text{lag}=L$. Sea $\mathcal{H}_{L+1}(w)$ la matriz de trayectorias tipo hankel por bloques de w , podemos expresar una representación del modelo \mathfrak{S} de la forma*

$$\tilde{\mathcal{B}} := \left\{ R \in \widehat{\mathcal{U}} : R\mathcal{H}_{L+1}(w) = 0 \right\} \quad (3.3)$$

Problema 5 (Identificación aproximada). *Una representación alternativa aproximada del Problema 4 se puede escribir de la forma*

$$\tilde{\mathcal{B}} := \left\{ R \in \mathbb{R}^{n \times (L+1)(m+n)} := \arg \min_{\hat{R}} \left\| \hat{R}\mathcal{H}_{L+1}(w) \right\|^2 \right\} \quad (3.4)$$

El algoritmo SpLra, para resolver el Problema 1, se puede resumir en los siguientes pasos:

1. Construir la matriz de trayectorias tipo hankel por bloques $\mathcal{H}_{L+1}(w)$.
2. Aplicar el Algoritmo 1 (lra) a $\mathcal{H}_{L+1}(w)$ para calcular

$$D(\mathcal{H}_{L+1}(w)) = \text{lra}(\mathcal{H}_{L+1}(w), r),$$

donde $r = (L + 1)(m + n) - n$.

3. Hacer la partición en bloques

$$D(\mathcal{H}_{L+1}(w)) = \begin{bmatrix} D_0 \\ D_1 \end{bmatrix}$$

,

donde $D_0 \in \mathbb{R}^{r \times N}$ y $D_1 \in \mathbb{R}^{n \times N}$,

4. Calcular x_{sp} , la solución esparcida del sistema $D_1 = x_{sp}D_0$

$$x_{sp} = \text{lsspsolver}(D_0, D_1, \epsilon, \delta)$$

5. Hacer $R_{sp} = \begin{bmatrix} -x_{sp} & I_n \end{bmatrix}$.

Algoritmo 3: SpLra.

Datos: $w \in \mathbb{R}^{(m+n) \times N}$, $\delta > 0$, $L \in \mathbb{Z}^+$, $\epsilon > 0$

Resultado: $R_{sp} = \text{SpLra}(w, L, \delta, \epsilon)$

$\mathcal{H}_{L+1}(w) \leftarrow \text{blkhank}(w, L+1)$, matriz de trayectorias tipo hankel en bloques.

$r \leftarrow (L+1)(m+n) - n$

$D \leftarrow \text{lra}(\mathcal{H}_{L+1}(w), r, \delta)$, aproximación de bajo rango

$$D_0 \in \mathbb{R}^{r \times N}, D_1 \in \mathbb{R}^{n \times N} \leftarrow D = \begin{bmatrix} D_0 \\ D_1 \end{bmatrix}$$

$x_{sp} \leftarrow \text{lsspsolver}(D_0, D_1, \delta, N, \epsilon)$

$$R_{sp} \leftarrow \begin{bmatrix} x_{sp} & -I_w \end{bmatrix}$$

3.4. Modelos Para Series de Tiempo

La econometría ha inspirado el desarrollo de modelos de series de tiempo y ecuaciones en diferencias. Por ejemplo los primeros trabajos en series de tiempo en [41, 39], el desarrollo de una teoría asintótica para el estimador de mínimos cuadrados (LS) para ecuaciones lineales estocásticas en diferencias presentada en [22] y su posterior extensión a sistemas simultáneos, donde LS no es consistente, desarrollada en [21].

Los modelos autorregresivos son ampliamente utilizados en muchos campos de las ciencias e ingeniería, por ejemplo en la predicción de precios de acciones financieras [11], en la predicción de la dinámica de la vegetación utilizando datos satelitales remotos [27], entre otros.

Un *proceso autorregresivo* de orden p , el cual en lo sucesivo denotaremos por

AR(p), es un proceso donde el valor de una serie de tiempo en el tiempo t , y_t , es la suma ponderada de los p valores pasados más un término de ruido denotado por ε_t . El proceso lo podemos representar por la siguiente ecuación en diferencias

$$y_t = a_1y_{t-1} + a_2y_{t-2} + \dots + a_py_{t-p} + \varepsilon_t, \quad (3.5)$$

donde $\varepsilon_t \sim N(0, \sigma^2)$. Si utilizamos el operador de retardo *lag*, denotado por L , el cual aplicado a la variable en el tiempo t se obtiene el valor de la variable en el tiempo $t - 1$, esto es $Ly_t = y_{t-1}$, el proceso (3.5) lo podemos expresar de la forma

$$(1 - a_1L - a_2L^2 - \dots - a_pL^p)y_t = \varepsilon_t$$

o de forma compacta

$$a(L)y_t = \varepsilon_t. \quad (3.6)$$

donde el polinomio autorregresivo $a(L)$ está definido por $a(L) = 1 - a_1L - a_2L^2 - \dots - a_pL^p$. Para estimar los coeficientes $a_i, i = 1, \dots, n$, podemos utilizar el método de mínimos cuadrados haciendo el cuadrado del error alcanzar su mínimo [40].

Un proceso de *medias móviles* de orden p , el cual de forma abreviada denotaremos por MA(p), es un proceso de sumas ponderadas de q elementos autorregresivos de disturbio más un término contemporáneo de disturbio, de la forma

$$y(t) = b_0\varepsilon_t + b_1\varepsilon_{t-1} + \dots + b_q\varepsilon_{t-q}, \quad (3.7)$$

lo cual se puede representar en forma abreviada

$$y_t = b(L)\varepsilon_t, \quad (3.8)$$

donde $\varepsilon_t \sim N(0, \sigma^2)$, b_0, \dots, b_q son constantes y $b(L) = b_0 + b_1L + \dots + b_qL^q$ representa el polinomio de medias móviles de grado q con $b_0 \neq 0$. Los procesos AR y MA presentados anteriormente se pueden unir para generar el proceso *autorregresivo de medias móviles*, el cual denotaremos por $\text{ARMA}(p, q)$ y estará representado por

$$y_t = a_1y_{t-1} + a_2y_{t-2} + \dots + a_py_{t-p} + \varepsilon_t + b_1\varepsilon_{t-1} + \dots + b_q\varepsilon_{t-q} \quad (3.9)$$

o de forma compacta

$$a(L)y_t = b(L)\varepsilon_t. \quad (3.10)$$

Es posible identificar un modelo $\text{ARMA}(p, q)$ utilizando una técnica autorregresiva de dos pasos. El método funciona de la siguiente manera:

1. Se inicia realizando una identificación de un modelo autorregresivo de orden m para y_t , es decir encontramos $\tilde{y}_t = \sum_{j=1}^m A'_j y_{t-j}$. Luego definimos \hat{u}_t de la forma

$$\hat{u}_t = y_t - \tilde{y}_t,$$

lo cual representa una estimación de los residuos u_t .

2. Utilizando la estimación encontrada para \hat{u}_t se procede a realizar una identificación del modelo autorregresivo para y_t de la forma

$$y_t = \sum_{j=1}^p A_j y_{t-j} + \sum_{j=1}^q A'_j \hat{u}_{t-j}.$$

Los procesos ARMA pueden dar un modelo reducido con relativamente menos parámetros desconocidos. Nos dan una alternativa a los modelos AR y MA de alto orden utilizando una combinación de ambos procesos. Si se incluyen variables determinadas de forma exógena, esto es que su valor está determinado por factores externos al modelo, se conoce como proceso ARMA con variables exógenas y lo denotaremos por ARMAX. Si utilizamos la variable x_t para la variable exógena, el proceso ARMAX lo podemos representar de la forma

$$a(L)y_t = b(L)\varepsilon_t + g(L)x_t. \quad (3.11)$$

El grado del polinomio $g(L)$ nos proporciona el número de elementos pasados de la variable exógena x_t , que influyen en la variable endógena y_t , es decir que su valor está determinado por factores dentro del modelo. Una extensión multivariada del modelo AR es el proceso autorregresivo vectorial VAR, el cual tiene la siguiente representación general

$$\mathbf{x}_t = \mathbf{A}_1\mathbf{x}_{t-1} + \mathbf{A}_2\mathbf{x}_{t-2} + \dots + \mathbf{A}_p\mathbf{x}_{t-p} + \mathbf{s}_t + \varepsilon_t, \quad (3.12)$$

donde $\mathbf{x}_t = (x_{1,t}, \dots, x_{n,t})^T$ es una serie de tiempo multivariada; A_i , $i = 1, \dots, n$ son matrices de $n \times n$; $\varepsilon_t = (\varepsilon_{1,t}, \dots, \varepsilon_{n,t})^T$ es ruido blanco multivariado con matriz de covarianza Ω y $\mathbf{s}_t = (s_{1,t}, \dots, s_{n,t})^T$ es un vector de términos determinísticos. En términos del operador L , un modelo VAR se puede escribir de la forma

$$\mathbf{x}_t = (A_1L + A_2L^2 + \dots + A_pL^p)\mathbf{x}_t + \mathbf{s}_t + \varepsilon_t.$$

En la mayoría de las aplicaciones, el término determinista \mathbf{s}_t consiste en términos constantes, por ejemplo $\mathbf{s}_t = \mathbf{v}$, donde \mathbf{v}_t es un vector constante. Un modelo de series de tiempo univariado está limitado por solo una parte de toda la información y por tanto una extensión natural son los modelos VAR. Sin embargo, en el caso que

se tenga disponible mucha información en series de tiempo, no resulta conveniente incluir todas las variables en el modelo. Tratar de estimar más parámetros puede llevar a resultados imprecisos, especialmente en los modelos VAR, donde el número de parámetros a estimar crece de manera cuadrática con respecto al número de variables.

Los vectores autorregresivos VAR, introducidos por Christopher Sims en [29], definieron un nuevo marco macroeconómico muy prometedor. Un modelo VAR es un modelo lineal donde el valor actual de una variable es proporcionado por sus valores anteriores más los valores pasados de las $n - 1$ variables restantes. Este sencillo proceso permitió una mejor visión de la dinámica en series de tiempo y además las herramientas estadísticas que conllevan los modelos VAR son fáciles de usar e interpretar.

Un modelo de series de tiempo univariado está limitado por solo incluir una parte de toda la información y por tanto una extensión natural son los modelos VAR. Sin embargo, en el caso que se tenga disponible mucha información en series de tiempo, no resulta conveniente incluir todas las variables en el modelo. Tratar de estimar más parámetros puede llevar a resultados imprecisos, especialmente en los modelos VAR, donde el número de parámetros a estimar crece de manera cuadrática con respecto al número de variables.

Una desventaja de los modelos VAR es el rápido crecimiento del número de los parámetros, con respecto al número de variables incluidas. Una alternativa para incluir más información en el modelo utilizando técnicas de reducción de orden, el modelo Factorial Dinámico Aproximado (DFM, por sus siglas en inglés) presentado en [31], el cual nos permite hacer predicciones de variables macroeconómicas de alta dimensionalidad.

En esta sección se presentan algunas técnicas utilizadas para modelar y pronosticar series de tiempo. El pronóstico de series de tiempo es una de las áreas más importantes en el estudio de variables económicas [42], donde observaciones pasadas

de una o varias variables son recolectadas y analizadas para desarrollar modelos que describan una relación subyacente. Para estimar los parámetros del modelo se utilizará un método de identificación esparcida por mínimos cuadrados planteado en [33].

3.5. Reducción de Dimensión y Modelos Factoriales Dinámicos

En el marco de nuestra investigación, hemos identificado una conexión fundamental entre los modelos factoriales dinámicos (DFM, por sus siglas en inglés) y la aproximación esparcida y regularizada de bajo rango. Los DFM son una clase de modelos estadísticos que se utilizan para capturar la estructura latente (estática) y la dinámica temporal en datos multidimensionales.

Problema 6. *Dada una secuencia de datos $Y = \{y_t\}_{t=1}^N \in \mathbb{R}^{n \times N}$ correspondiente a una serie de tiempo $\{y_t : t \geq 1\} \subset \mathbb{R}^n$, los enteros r, L tal que $0 < r < n$ y $0 < L \ll N$, encontrar*

$$\begin{bmatrix} \mathcal{L} & \mathcal{F} \end{bmatrix} = \arg \min_{\hat{\mathcal{L}}, \hat{\mathcal{F}}} \left\| Y - \hat{\mathcal{L}} \hat{\mathcal{F}} \right\|_F^2$$

sujeto a: $\text{rank}(Y) \leq r$.

Luego construir un modelo autorregresivo de orden L , para la serie $\mathcal{F} = \{f_t\}_{t=1}^N \in \mathbb{R}^{r \times N}$, tal que la estructura general del modelo se pueda expresar de la forma definida en 2.5 y 2.6.

Los siguientes pasos propuestos nos permiten dar una solución al Problema 6:

1. Aplicar la descomposición truncada en valores singulares SVD de la matriz original de datos Y de la misma forma que en el Teorema 2 utilizando el

Algoritmo 1.

$$Y_r = U_r \Sigma_r V_r^T$$

donde $U_r \in \mathbb{R}^{r \times n}$, $\Sigma_r \in \mathbb{R}^{r \times r}$ y $V_r \in \mathbb{R}^{r \times N}$.

2. Hacer $\mathcal{L} = U_r \sqrt{\Sigma_r}$ y $\mathcal{F} = \sqrt{\Sigma_r} V_r^T$.
3. Aplicar **SpLra** a \mathcal{F} .

En forma de algoritmo:

Algoritmo 4: `SpLra_dfm`

Datos: $Y \in \mathbb{R}^{n \times N}$, $r \in \mathbb{Z}^+$, $L \in \mathbb{Z}^+$, $\delta > 0$, $\epsilon > 0$

Resultado: $R_{sp}, \mathcal{L}, \mathcal{F} = \mathbf{SpLra_dmf}(Y, r, L, \delta, \epsilon)$

$U_r, S_r, V_r \leftarrow \mathbf{lra}(Y, r)$, aproximación de bajo con el Algoritmo 1.

$\mathcal{L} \leftarrow U_r \sqrt{S_r}$

$\mathcal{F} \leftarrow \sqrt{S_r} V_r^T$

$R \leftarrow \mathbf{SpLra}(\mathcal{F}, L, \delta, \epsilon)$

Los modelos factoriales dinámicos, en nuestra interpretación anterior, desempeñan un papel esencial en la reducción de la dimensión de los datos. Su aplicación nos proporciona una herramienta poderosa para abordar nuestros objetivos de investigación y extraer conocimientos significativos.

La conexión entre los DFM y nuestra investigación se basa en la capacidad de estos para modelar la complejidad de nuestros datos y descomponerlos en componentes interpretables por medio de la aproximación de bajo rango, lo que enriquece nuestro análisis y conclusiones.

Capítulo 4

Experimentos y Resultados

En el proceso de investigación presentado en este documento, es fundamental garantizar la aplicabilidad de las técnicas propuestas. Para lograr este objetivo, se han diseñado y ejecutado un conjunto de experimentos completamente replicables, que abarcan tanto datos simulados como datos reales.

Para evaluar la efectividad y el rendimiento de las técnicas teóricas y computacionales propuestas, comenzamos utilizando datos simulados. Estos datos fueron generados sintéticamente, con el fin de simular situaciones de interés en nuestra investigación. Esta fase nos permitió establecer una línea de partida y comprender cómo se comportan nuestras técnicas en condiciones ideales y controladas.

Reconociendo la importancia de la aplicabilidad en el mundo real, también realizamos experimentos utilizando datos reales. Estos datos se obtuvieron de bases de datos públicas, en diversas áreas relacionadas del conocimiento. Esta parte nos permitió evaluar la eficiencia de nuestras técnicas en situaciones reales, donde los desafíos y la complejidad pueden ser más significativos.

Preprocesamiento de los datos

Los datos recopilados se sometieron a un proceso de limpieza y preprocesamiento. En la sección de cada experimento se detalla cada uno de los métodos utilizados para garantizar la coherencia y la calidad de los datos utilizados.

A continuación algunos métodos utilizados:

1. Estandarización para llevar todos los datos a una escala común.
2. Eliminación de tendencia, utilizando técnicas de regresión. Esto implica ajustar una línea de regresión o una curva a los datos que representa la tendencia, y luego sustraer esta tendencia del conjunto de datos original.
3. Separación del conjunto de datos en conjuntos de identificación y validación.

Métricas de Evaluación

- Porcentaje de ajuste

$$\%fit = 100 \left(1 - \frac{\|w - w_{id}\|}{\|w - \bar{w}\|} \right) \quad (4.1)$$

donde w es la serie de tiempo original, \bar{w} es su valor promedio y w_{id} es la serie identificada.

- El $rmse$ mide la diferencia entre los valores reales observados y los valores predichos por el modelo, calculando la raíz cuadrada de la media de los errores al cuadrado.

$$rmse = \sqrt{\frac{1}{N} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (4.2)$$

donde y_i representa el valor real u observado, \hat{y}_i el valor predicho y N el total de observaciones.

- Coeficiente de determinación, es una métrica estadística utilizada para evaluar la bondad de ajuste de un modelo.

$$R^2 = 1 - \frac{SSR}{SST} \quad (4.3)$$

donde SSR es la suma de los cuadrados de las diferencias entre los valores predichos por el modelo y la media de la variable dependiente y SST es la suma de los cuadrados de las diferencias entre los valores reales y la media de la variable dependiente.

Disponibilidad de los Datos y Código

Todos los datos, archivos y la implementación de los algoritmos y experimentos están disponibles para ser ejecutados y reproducidos en [12].

4.1. Identificación de un Sistema SISO

El siguiente experimento está basado en algunas ideas presentados en [2], los datos del proceso de identificación son simulados utilizando un sistema de una entrada y una salida (SISO por sus siglas en inglés) con la siguiente función de transferencia

$$G(z) = \frac{1.5z^2 - 2.07z + 1.315}{z^{14} - 2.21z^{13} + 1.7497z^{12} - 0.5843z^{11} + 0.0684z^{11}} \quad (4.4)$$

la señal de entrada es una secuencia pseudo-aleatoria binaria (PRBS por sus siglas en inglés), con una probabilidad de cambio de 8 % y un rango entre $[-1, 1]$. La figura 4.1 muestra la salida gráfica de las señales de entrada y salida u_t, y_t respectivamente.

Con el objetivo de validar los diferentes modelos identificados, se generan nuevos datos a partir de otra señal PRBS. Estos datos generados no participan en el proceso de identificación. La salida gráfica para los datos del proceso de validación se presentan en la Figura 4.3.

A continuación se procede a perturbar el sistema con una estructura ARMAX,

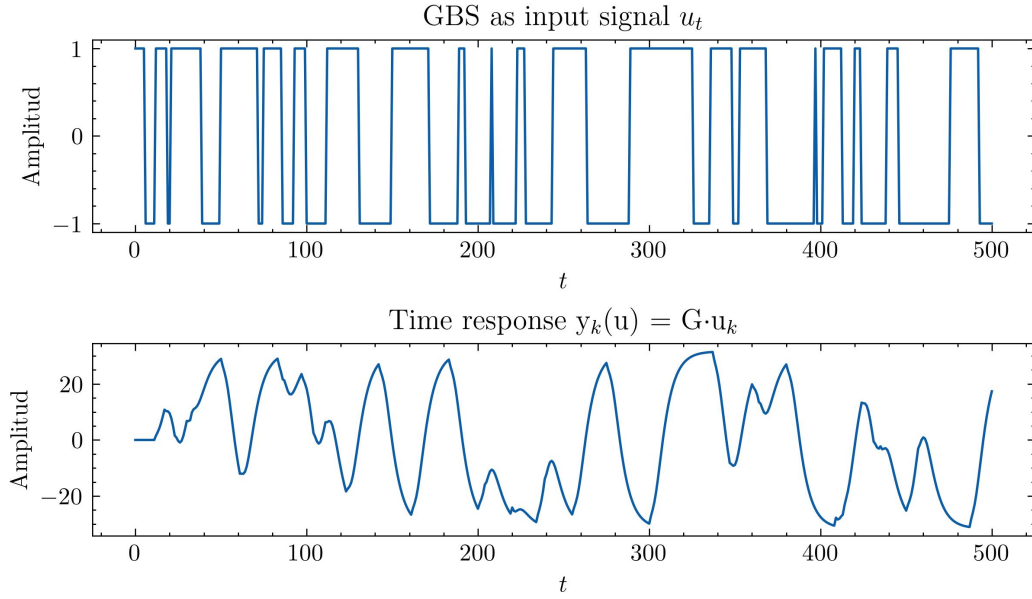


Figura 4.1: Serie de tiempo para la fase de identificación.

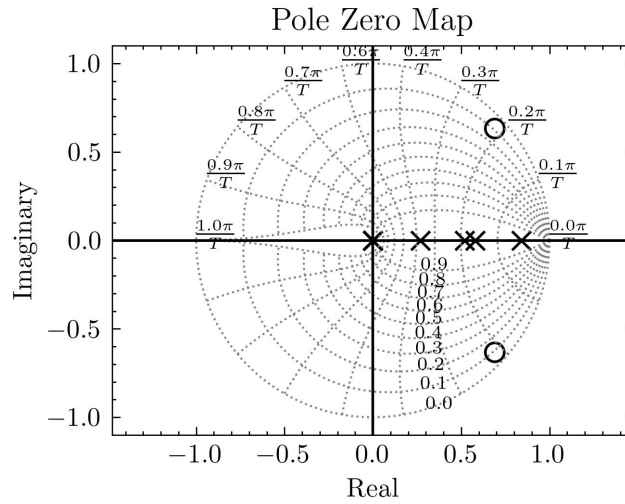


Figura 4.2: Polos y ceros del sistema simulado original.

generando cuatro señales independientes de ruido blanco, $\sigma_1^2 = 0.000$, $\sigma_2^2 = 0.001$, $\sigma_3^2 = 0.010$ y $\sigma_4^2 = 0.1000$ filtradas por la siguiente función de transferencia

$$H(z) = \frac{z^{14} + 0.3z^{13} + 0.2z^{12}}{z^{14} - 2.21z^{13} + 1.7497z^{12} - 0.5843z^{11} + 0.0684z^{11}} \quad (4.5)$$

Tres tipos de modelos son identificados, los modelos ARX y ARMAX son calculados utilizando el paquete `System Identification Package for PYthon (SIPPY)` [2] y nuestro modelo propuesto `SpLra` con los algoritmos implementados con código

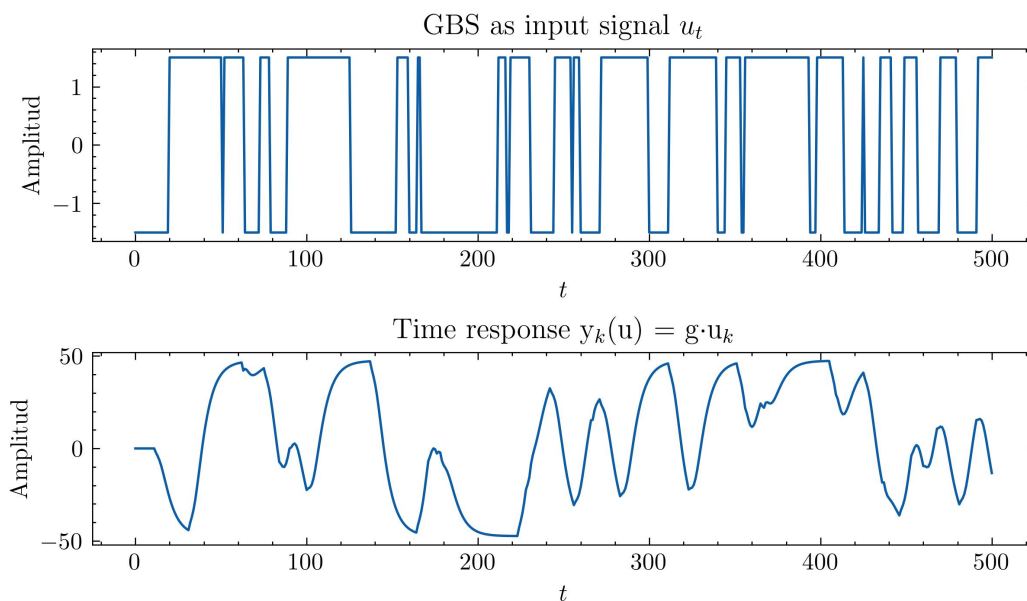


Figura 4.3: Serie de tiempo para la fase de validación.

Python en [12].

La dificultad de este experimento radica en seleccionar el orden adecuado del sistema, esto debido a un nivel alto de desfase en la función de transferencia $G(z)$. El paquete SIPPY cuenta con tres criterios de selección del orden de los modelos, **AIC**, **BIC** y **AcIC**, los cuales son los criterios mayormente utilizados en la práctica. En el método **SpLra** se utiliza un método gráfico que consiste en definir un número máximo para L y graficar los diferentes valores con respecto al desajuste producido por cada modelo.

La tabla 4.1 resume los resultados para los diferentes modelos identificados, donde **nnz (%)** es el porcentaje de elementos distintos de cero de los parámetros del modelo y **%fit** es el porcentaje de ajuste.

	Modelo					
	Arx		Armax		SpLra	
nnz	280		47		333	
σ^2	%fit(id)	%fit(val)	%fit(id)	%fit(val)	%fit(id)	%fit(val)
0.000	81.97	80.67	81.96	80.67	99.99	99.99
0.001	81.81	80.47	$-\infty$	$-\infty$	99.28	99.33
0.010	83.13	82.12	$-\infty$	$-\infty$	96.55	96.37
0.100	78.05	77.47	$-\infty$	$-\infty$	91.81	92.64

Tabla 4.1: Resumen de resultados.

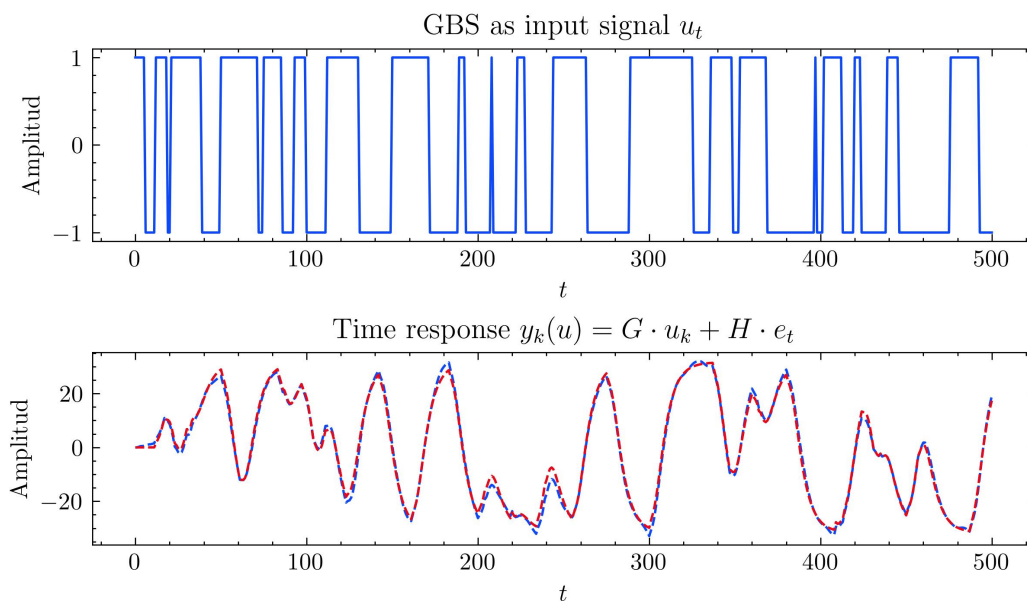


Figura 4.4: Serie de tiempo del sistema perturbado con una estructura ARMAX con $\sigma^2 = 0.010$.

Los resultados muestran que el método propuesto **SpLra** obtuvo mejores resultados en el ajuste del modelo con un menor número de parámetros. Es importante aclarar que los resultados de los modelos **ARX** y **ARMAX** se pueden mejorar afinando el orden, lo cual requiere pasos adicionales. Notar que el modelo **ARMAX** presenta valores de $\%fit = -\infty$, lo cual indica que el modelo no es estable. La idea del experimento es mostrar como el enfoque integrado utilizado en **SpLra** facilita muchos aspectos del proceso de identificación.

4.2. Identificación de un Sistema MIMO

El presente experimento está parcialmente motivado por los resultados presentados en [2]. Los datos son generados simulando un sistema MIMO (múltiples entradas y múltiples salidas, por sus siglas en inglés), con $m = 4$ entradas y $p = 3$ salidas. Los diferentes ordenes y especificaciones de la función de transferencia se muestran en la Tabla 4.2

Como señales de entrada se generan de forma independientes cuatro secuencias pseudo-aleatorias PRBS, con probabilidad de cambio de 8% y un rango entre $[-1, 1]$.

	Salida 1	Salida 2	Salida 3
Entrada 1	$\frac{0.4z^3+0.33z^2}{z^5-0.3z^4-0.25z^3-0.021z^2}$	$\frac{-0.85z^2-0.575z-0.277}{z^5-0.4z^4}$	$\frac{0.2z^3}{z^5-0.1z^4-0.3z^3}$
Entrada 2	$\frac{0.10z^2}{z^5-0.3z^4-0.25z^3-0.021z^2}$	$\frac{0.71z-0.123}{z^5-0.4z^4}$	$\frac{0.821z^2+0.423z}{z^5-0.1z^4-0.3z^3}$
Entrada 3	$\frac{0.7z^2+0.75z+0.22}{z^5-0.3z^4-0.25z^3-0.021z^2}$	$\frac{-0.1z^3}{z^5-0.4z^4}$	$\frac{-0.1z^3}{z^5-0.1z^4-0.3z^3}$
Entrada 4	$\frac{-0.9z^3-0.11z^2}{z^5-0.3z^4-0.25z^3-0.021z^2}$	$\frac{-0.994z^3}{z^5-0.4z^4}$	$\frac{0.891z+0.223}{z^5-0.1z^4-0.3z^3}$

Tabla 4.2: Función de transferencia sistema MIMO original.

La salida gráfica de estas señales de entrada y salida se muestran en la Figura 4.5.

Con el propósito de mostrar la consistencia de los métodos, se distorsionan las señales de salida solamente, con un ruido pseudo-aleatorio con diferentes niveles de varianza ($\sigma^2 = 0.000$, $\sigma^2 = 0.001$, $\sigma^2 = 0.010$ y $\sigma^2 = 0.100$).

Como modelo comparativo se utilizó el método N4SID (Non -iterative method for system identification) implementado en [2]. Dicho método es uno de las variantes de los métodos conocidos como *de subespacio*, los cuales devuelven los parámetros del modelo utilizando una representación de espacio de estados.

No se asume un orden conocido del modelo. En el caso de N4SID se utiliza el criterio AIC y en el caso de SpLra el método gráfico implementado en la función `lag_est` de [12]. Los resultados para este experimento se muestran en la Tabla 4.3.

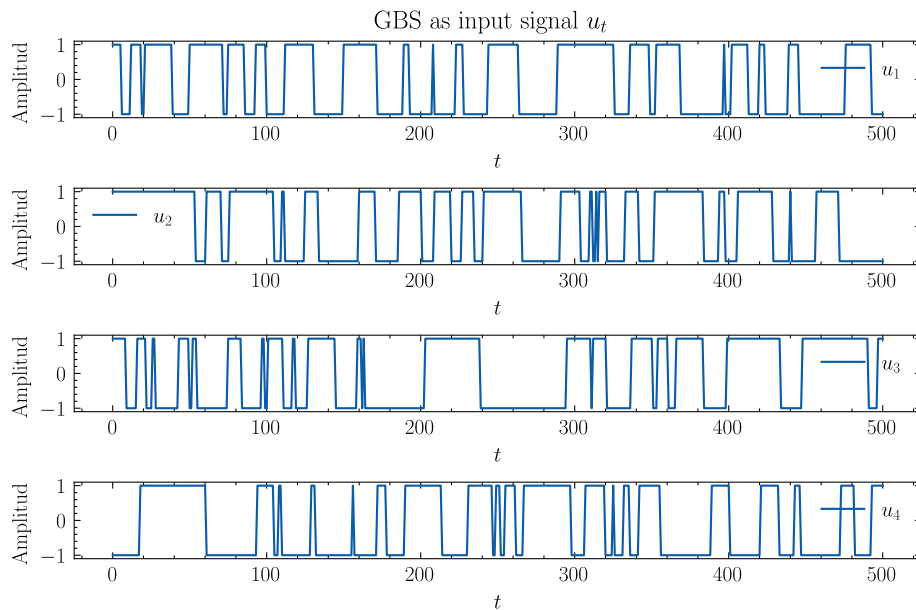


Figura 4.5: Señales de entrada PBRs.

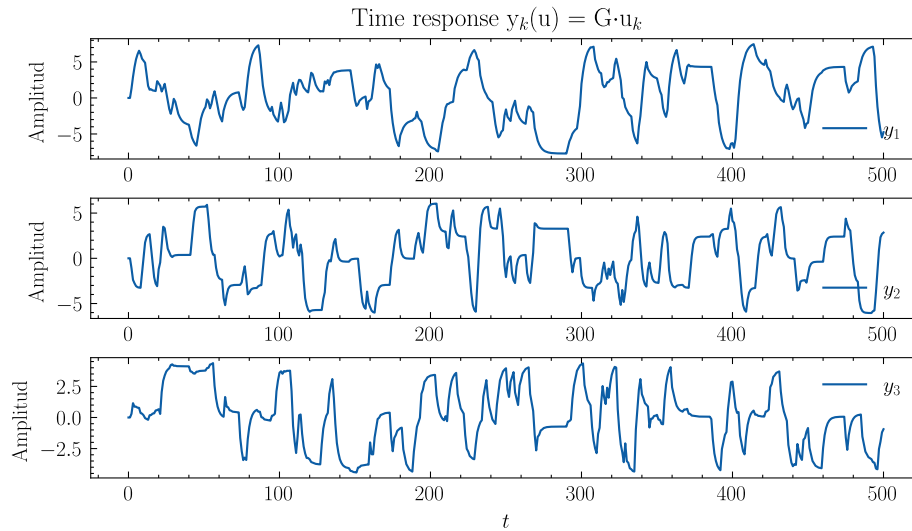


Figura 4.6: Señales de salida generadas a partir del sistema original.

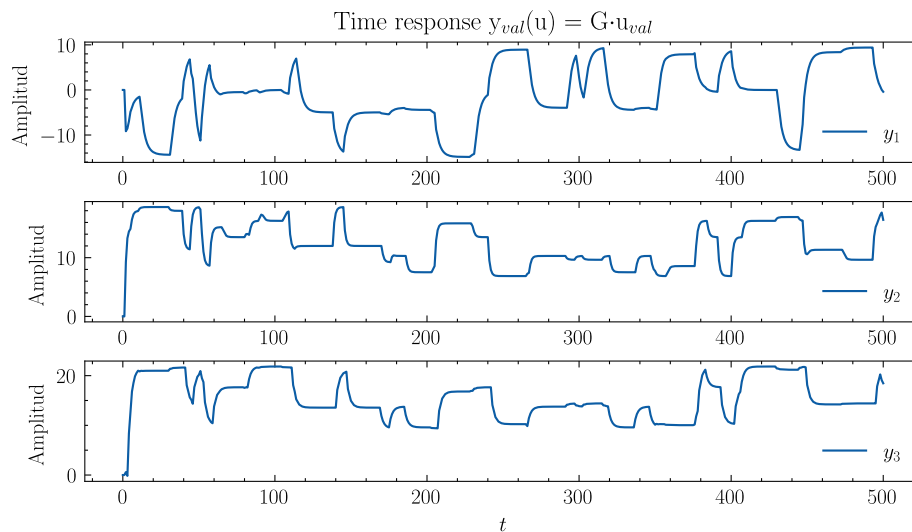


Figura 4.7: Señales de salida del sistema original a partir de un segundo grupo de señales de entrada. Estos datos simulados se utilizan en el proceso de validación.

nnz(%)	Modelo			
	N4SID		SpLra	
	280		333	
σ^2	%fit(id)	%fit(val)	%fit(id)	%fit(val)
0.000	97.14	97.66	99.99	99.99
0.001	97.13	97.63	99.94	99.98
0.010	97.06	97.33	99.13	99.27
0.100	93.01	96.37	93.63	97.35

Tabla 4.3: Resumen de resultados sistema MIMO.

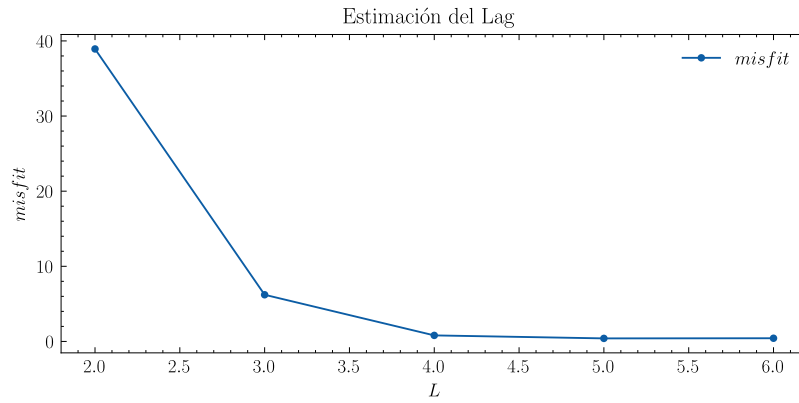


Figura 4.8: Salida gráfica del desajuste, para diferentes valores del orden del modelo. La grafica nos sugiere utilizar un valor de $L = 5$.

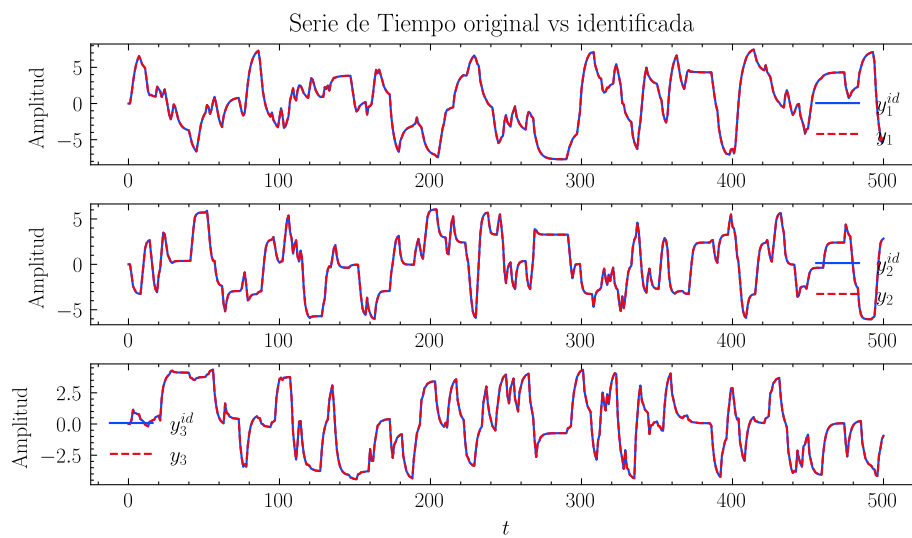


Figura 4.9: Señales de salida del sistema original e identificada.

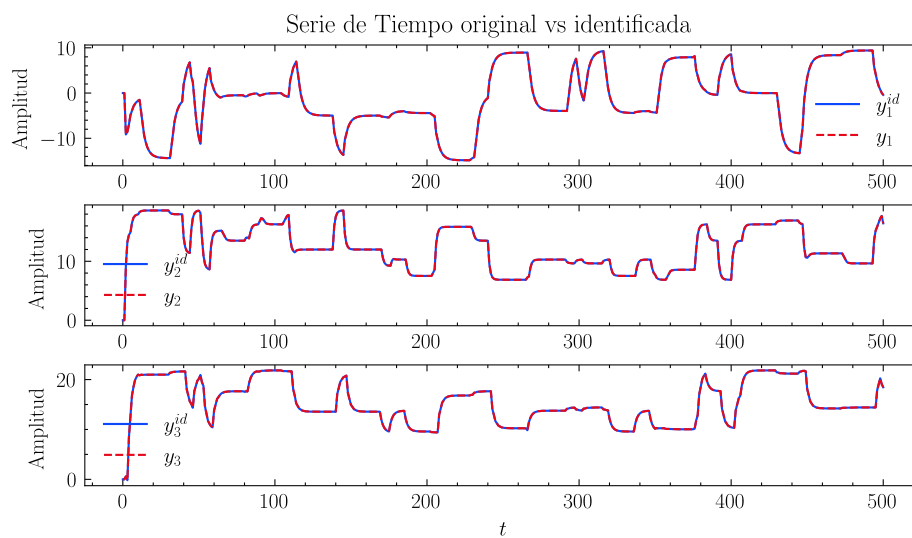


Figura 4.10: Señales de salida original e identificada en la fase de validación.

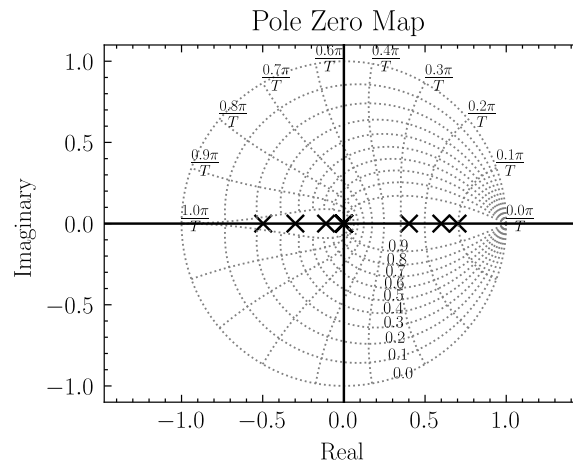


Figura 4.11: Salida gráfica de los ceros del sistema identificado.

4.3. Aplicaciones En Procesos Industriales

La base de datos para la identificación de sistemas DAISY [13] contiene diferentes conjuntos de datos reales en varias disciplinas, por ejemplo, sistemas de procesos industriales, sistemas eléctricos y electrónicos, sistemas mecánicos, entre otros. A continuación una breve descripción de los datos seleccionados.

Brazo robótico flexible: Estos datos son el resultado del movimiento de un brazo robótico conectado a un motor eléctrico. Los datos modelan la función de transferencia a partir de la reacción por el torque en la aceleración de la estructura. Los datos de entrada son el torque de reacción de la estructura y la salida la aceleración del brazo robótico.

Secador de pelo: Experimento de laboratorio que funciona como un secador de pelo. El aire es ventilado por medio de un tubo y calentado en su entrada. La temperatura del aire es medida cuando sale del tubo. Como entrada se mide el voltaje sobre la calefacción, la cual es una malla de cables de resistencia. La entradas es el voltaje del dispositivo de calentamiento y la salida la temperatura del aire de salida.

Reactor de un tanque: Este proceso modela el flujo continuo de un reactor de tanque de agitación, donde la reacción exotérmica y la concentración es regulado mediante un refrigerante. La entradas es el flujo del refrigerante (l/min) y las salidas son la concentración del flujo (mol/l) y la temperatura en (grados K).

Secador industrial: Datos de un secador industrial por Cambridge Control Ltd. Las entradas son la tasa de flujo de combustible, velocidad del extractor de gas caliente y tasa de flujo de materia prima. Como salidas la temperatura de bulbo seco, temperatura de bulbo húmedo y el contenido de humedad de la materia prima.

La Tabla 4.4 resume la cantidad de datos N , el número de variables de entrada m , cantidad de variables de salida n y el número de observaciones pasadas lag a considerar. La Tabla 4.5 resume los resultados obtenidos en este experimento. Para los datos 1 y 2 se consideraron los métodos **Arx** y **N4SID**, el último es un algoritmo de subespacio no iterativo implementado en el paquete **SIPPY** que calcula directamente una representación del modelo en variables de espacio de estados. Para los datos 3 y 4 se consideran los métodos **N4SID** y **MOESP**, el segundo, también implementado en el paquete **SIPPY** es una versión mejorada del primero.

No.	Nombre	N	m	n	lag
1	Brazo robótico flexible	1024	1	1	6
2	Secador de pelo	1000	1	1	5
3	Reactor de un tanque	7500	3	3	1
4	Secador industrial	867	3	3	1

Tabla 4.4: Selección de conjuntos de datos DAISY.

Resultados

1. Brazo robótico flexible

$$H_{id}(z) = \frac{-0.05z^6 - 0.31z^5 + 1.21z^4 - 1.84z^3 + 1.64z^2 - 0.77z + 0.12}{z^6 - 2.42z^5 + 2.13z^4 - 0.29z^3 - 0.46z^2 - 0.07z + 0.28}$$

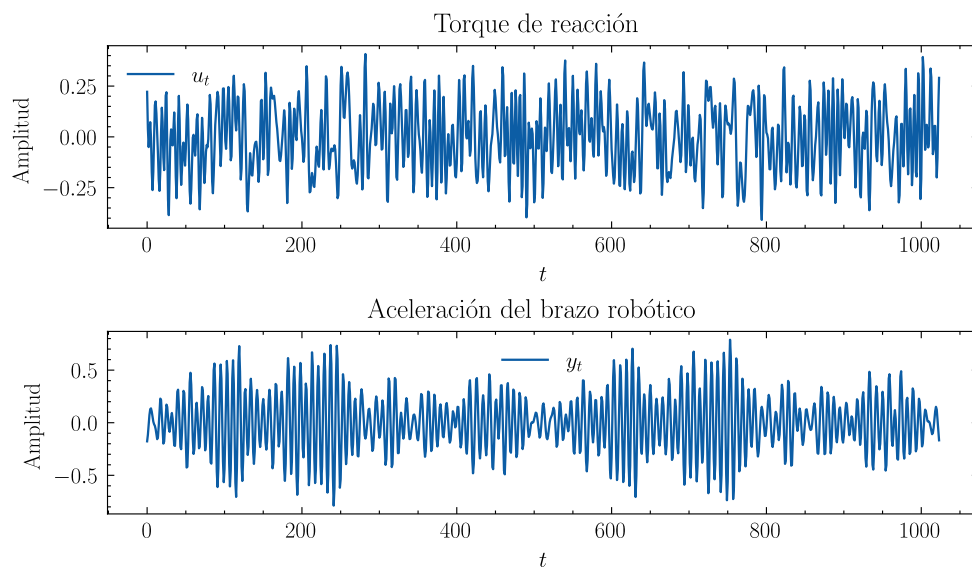


Figura 4.12: Salida gráfica de los datos de entrada y salida de un brazo robótico flexible .

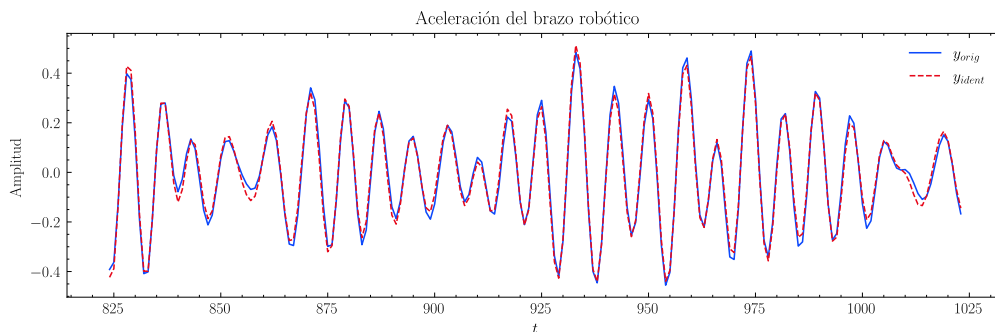


Figura 4.13: Salida gráfica de los últimos 200 datos de salida originales e identificados .

2. Secador de pelo

$$H_{id}(z) = \frac{0.066z^2 - 0.024z + 0.034}{z^5 - 2.32z^4 + 2.86z^3 - 2.49z^2 + 1.38z - 0.37}$$

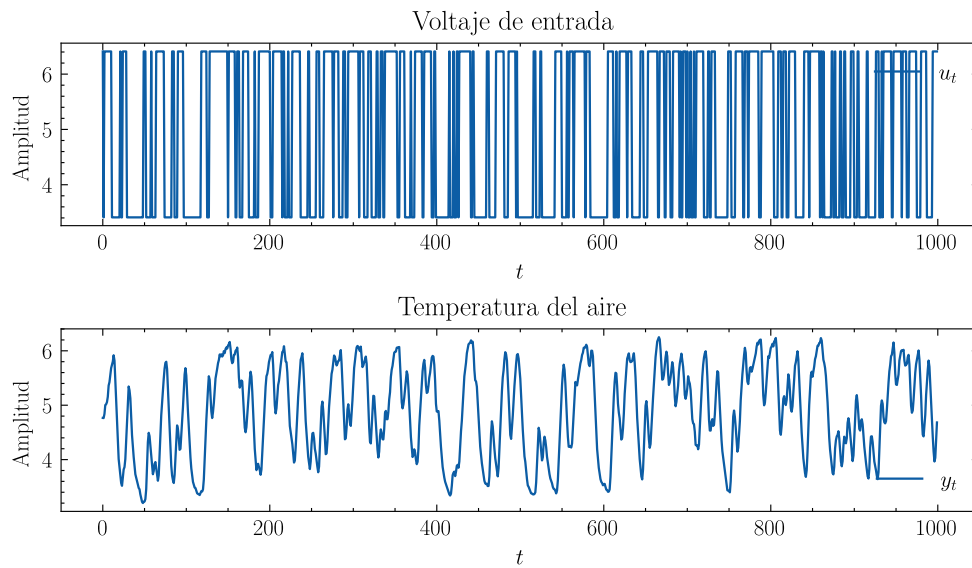


Figura 4.14: Salida gráfica de los datos de entrada y salida de un secador .

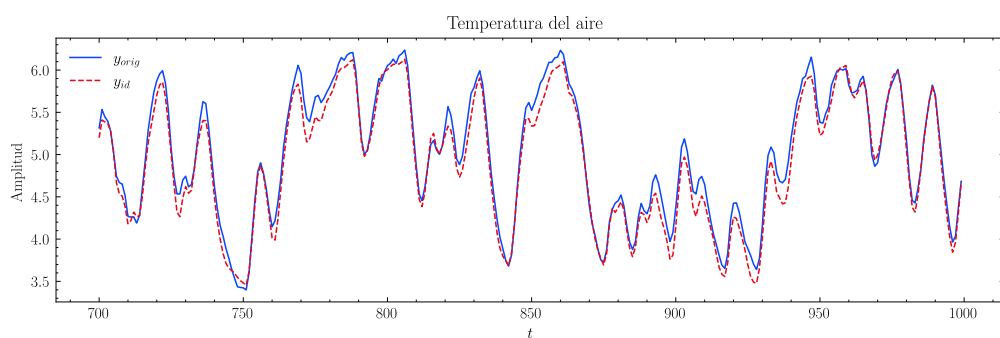


Figura 4.15: Salida gráfica de los últimos 300 datos de salida originales e identificados

3. Reactor de un tanque

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix} = \begin{bmatrix} 0.617 & -0.366 & | & 0.065 \\ 0.258 & 1.120 & | & -0.154 \\ \hline 1.000 & 0.000 & | & -0.012 \\ 0.000 & 1.000 & | & 0.000 \end{bmatrix},$$

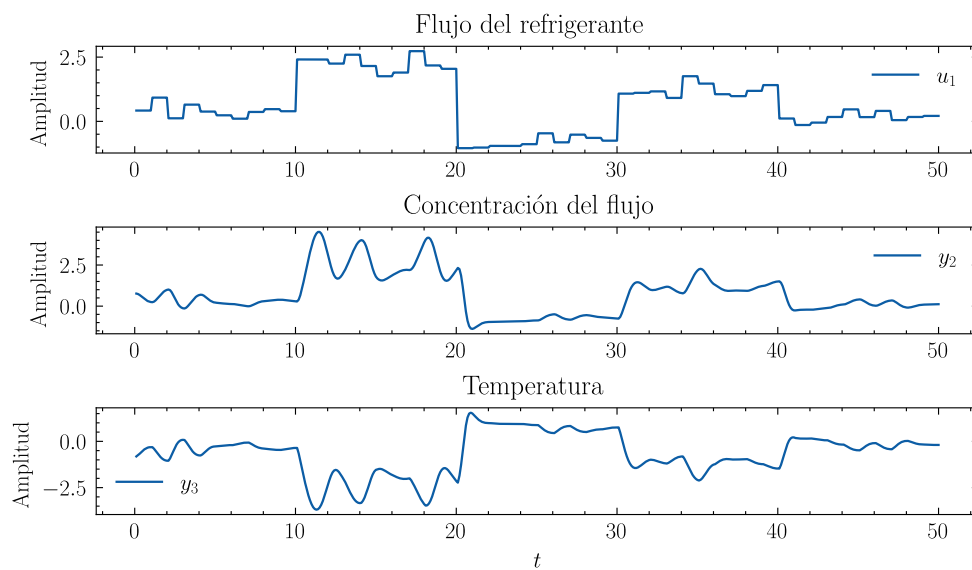


Figura 4.16: Salida gráfica de los primeros 500 datos(s) de salida originales .

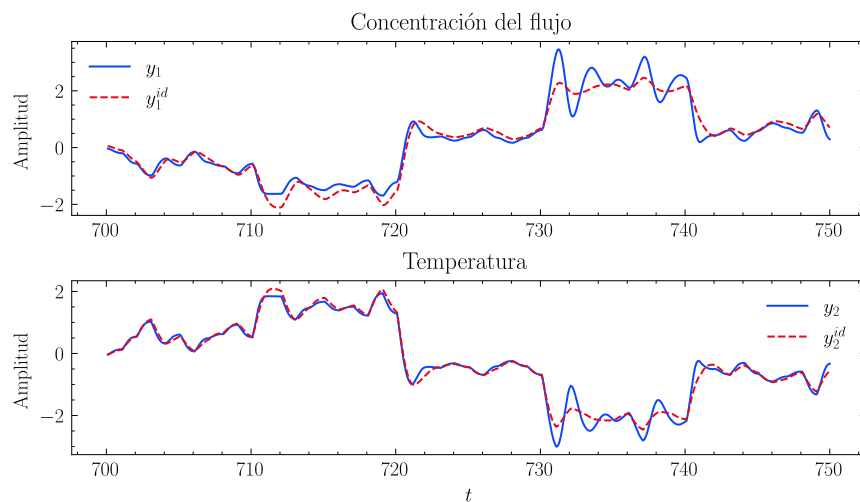


Figura 4.17: Salida gráfica de los últimos 500 datos(s) de salida originales e identificados .

No.	Nombre	Método	%fit(id)	%fit(val)	nnz(%)
1	Brazo robótico	Arx	85.27	82.39	100
		SpLra	96.99	92.64	38.78
		N4SID	81.60	91.48	97.96
2	Secador de pelo	Arx	87.45	82.39	66.66
		SpLra	93.34	92.24	36.11
		N4SID	87.59	82.11	97.22
3	Reactor de un tanque	SpLra	84.11	82.76	75.00
		N4SID	84.22	86.25	83.33
		MOESP	84.16	86.17	83.33
4	Secador industrial	SpLra	61.84	53.87	75.00
		N4SID	58.27	36.01	75.00
		MOESP	58.32	35.97	75.00

Tabla 4.5: Resultados base de datos Daisy.

4.4. Nivel de Agua Lago Erie

En esta sección se utilizan los datos proporcionados en [26], que corresponden a los niveles de agua del Lago Erie de 1921 a 1970. El estudio de las variaciones en los niveles del agua en los grandes lagos afecta directamente las economías locales, por tanto es importante construir modelos que permitan planificar actividades con mayor precisión. La salida gráfica para esta serie de tiempo se muestra en la Figura 4.18.

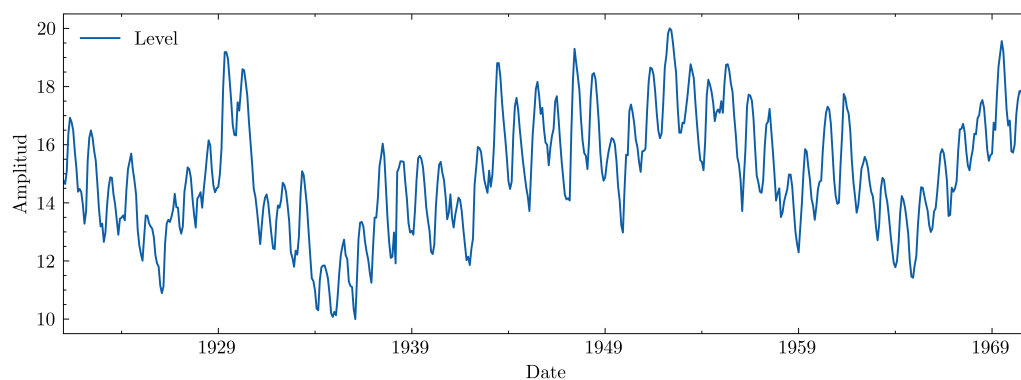


Figura 4.18: Niveles de agua Lago Erie, 1921-1970.

Como parte del pre-procesamiento de los datos se procede a eliminar tendencia lineal y estandarización. Luego se dividen los datos en conjuntos de identificación y validación, tal como lo muestra la Figura 4.19.

La estrategia de selección del orden del modelo L , consiste en calcular el $rmse$ para

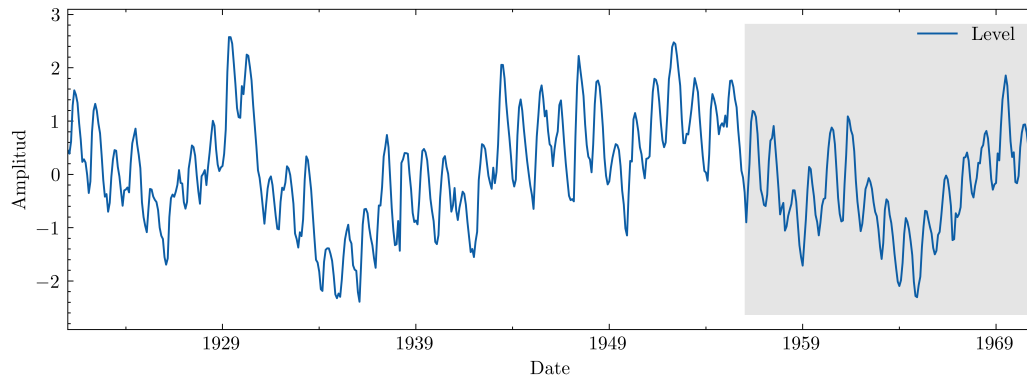


Figura 4.19: Salida gráfica de los datos sin tendencia y estandarizados. La parte sombreada representa el 30 % del total para el proceso de validación.

los conjuntos de datos de identificación y validación. Luego se puede considerar, de forma balanceada, un valor de L cuyos errores estén lo más cercano posible.

Observando la Figura 4.20 se decide tomar un valor de $L = 3$.

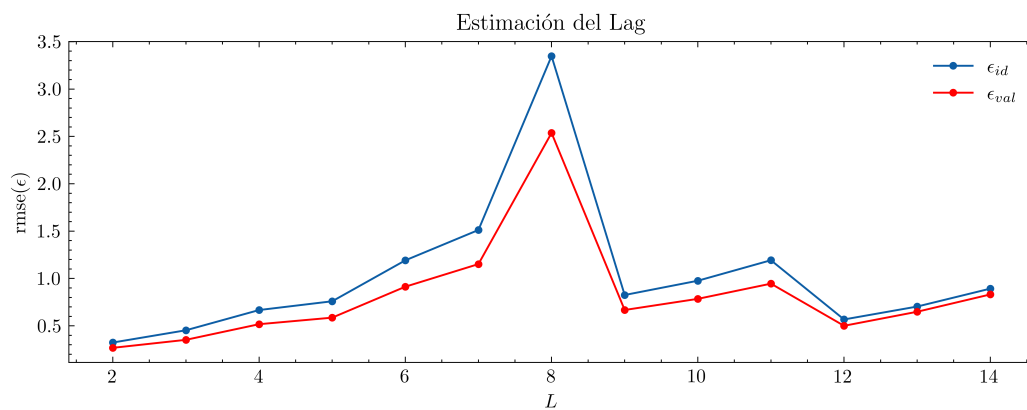


Figura 4.20: Comportamiento del $rmse$ de los datos de identificación y validación con respecto al orden ($L=lag$) del modelo.

A continuación se detallan los pasos propuestos en la Sección 3.4 para la identificación de sistemas ARMA:

- Estimación de la señal de errores ϵ_i
- Se construye la nueva serie de tiempo $w = \begin{bmatrix} \epsilon & y \end{bmatrix}^T$, donde y representa los datos originales y ϵ los datos de entrada del sistema. A continuación se presentan los sistemas identificados en representación de espacio de estados y como una función de transferencia.

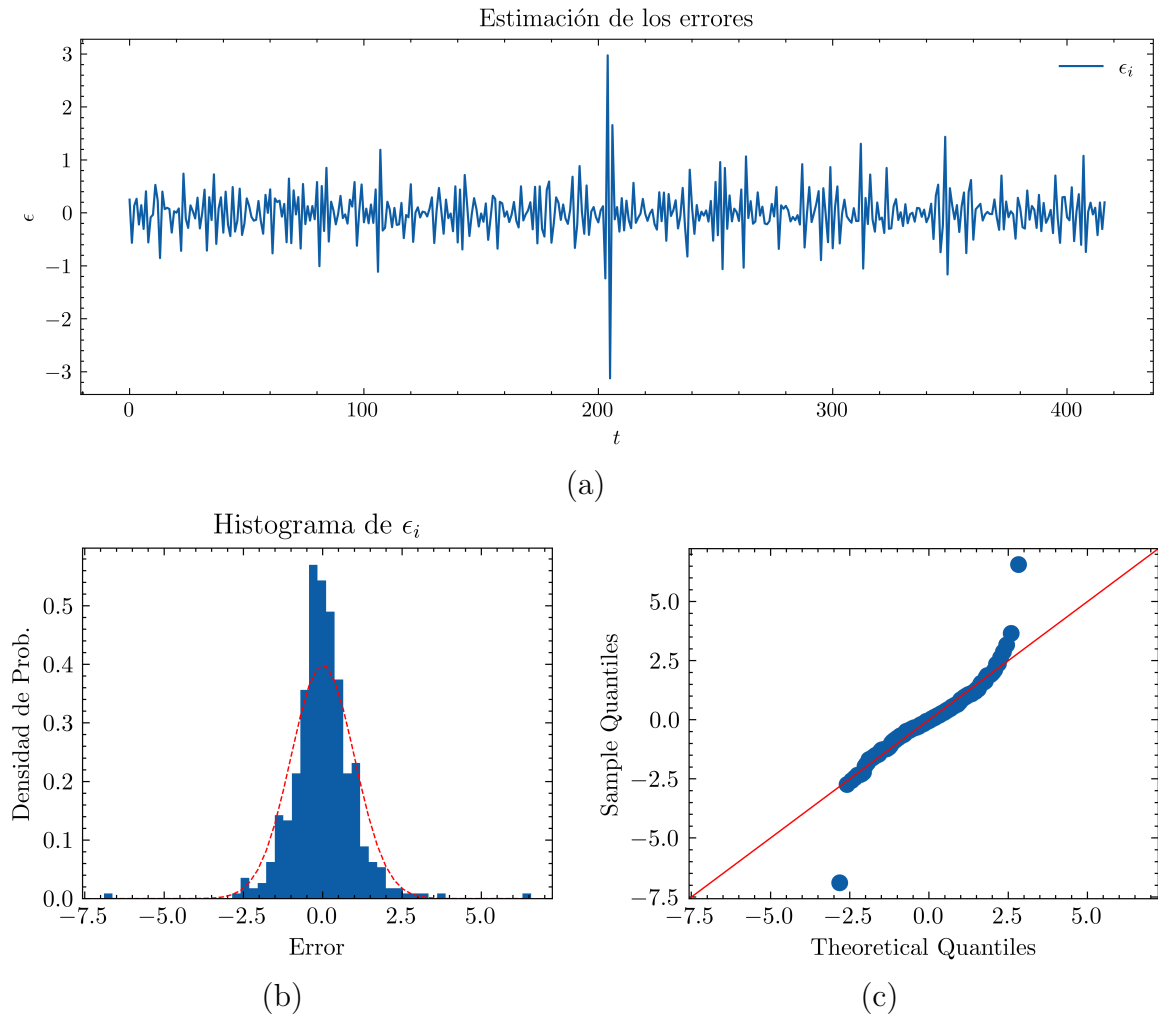


Figura 4.21: (a) Residuos del modelo autor-regresivo AR(3) con respecto a la serie original de datos. (b) Histograma de los residuos ϵ_i estandarizados. (c) Q-Q plot.

$$\left[\begin{array}{c|c} A & B \\ \hline C & D \end{array} \right] = \left[\begin{array}{ccc|c} 0 & 0 & 1 & 1 \\ 1 & 0 & -2.6 & -2.6 \\ 0 & 1 & 2.6 & 2.6 \\ \hline 0 & 1 & 2.6 & 1 \end{array} \right],$$

$$H = \frac{z^3}{z^3 - 2.598z^2 + 2.598z - 1}.$$

- Simulación y predicción

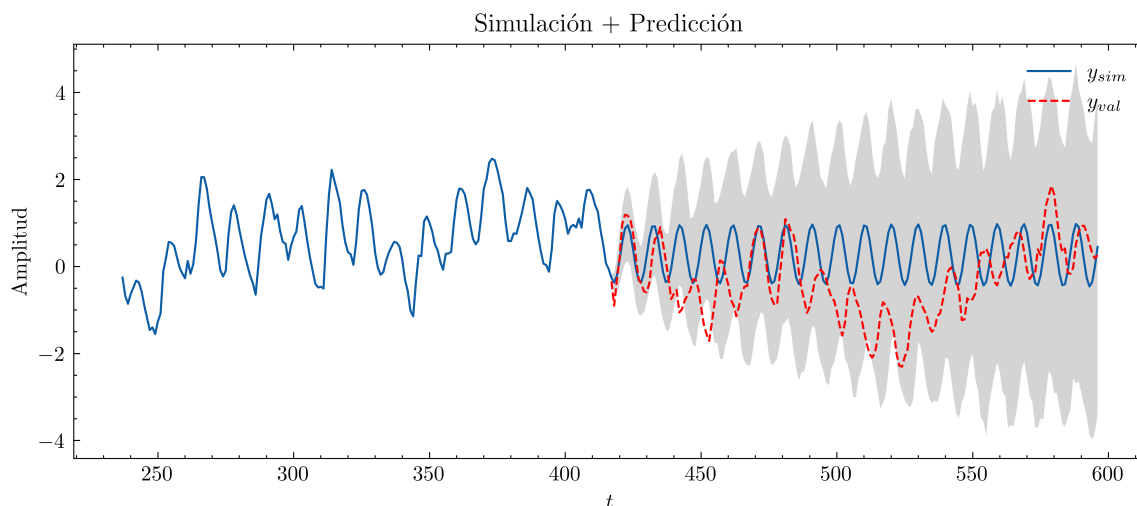


Figura 4.22: Niveles de agua Lago Erie, 1921-1970.

4.5. Aplicaciones En Econometría

La base de datos FRED-MD (Federal Reserve Economic Data - Macro Data, por sus siglas en inglés) [25], consiste en 127 series mensuales con indicadores macroeconómicos claves de los Estados Unidos desde el año 1959 hasta el 2021. La base de datos está diseñada para actualizarse mensualmente con acceso público gratuito. FRED-MD no solo ofrece una amplia gama de series temporales, sino que también se ha diseñado meticulosamente para simplificar la tarea de los investigadores al comparar y llevar a cabo análisis empíricos.

Entre las series disponibles se encuentran indicadores importantes como el Producto Interno Bruto, el desempleo, la inflación, las tasas de interés y muchas otras. La amplia cobertura temporal de la base de datos permite a los investigadores rastrear y analizar tendencias macroeconómicas a lo largo de más de seis décadas.

Los resultados obtenidos en esta sección, están basados parcialmente, en replicar algunas ideas presentadas en [9]. La motivación detrás de esta elección radica en la importancia de comprender y modelar adecuadamente la dinámica macroeconómica, especialmente en lo que respecta a la tasa de desempleo (UNRATE) como variable objetivo.

Con el objetivo de seleccionar un subconjunto de variables adecuado ($n \leq 10$) de

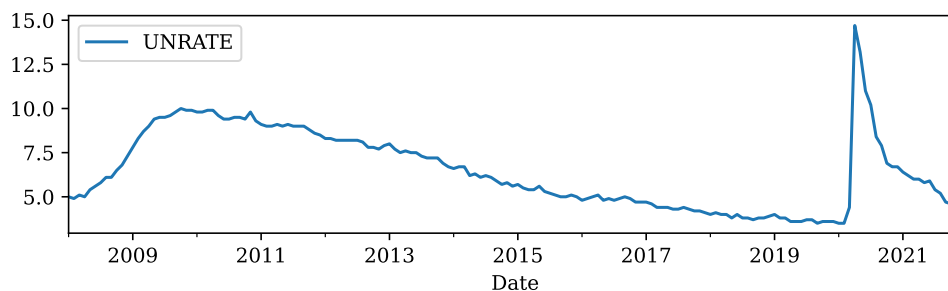


Figura 4.23: Serie de tiempo tasa de desempleo *UNRATE* en USA, 2008-2022.

la base de datos FRED-MD, tomando en cuenta el experimento realizado en [25], se extraen nueve series descritas en la Tabla 4.6.

Variable	Código FRED
Desempleo	UNRATE
Tasa de descuento, Bonos del Tesoro (3 meses)	TB3MS
Ingreso Real	RPI
Índice de Prod. Industrial	INDPRO
S&P 500	S&P 500
Préstamos Negocios	BUSLOANS
CPI	CPIAUCS
Precio crudo	OILPRICE _x
M2 Money	M2SL

Tabla 4.6: Selección de variables macroeconómicas.

La Figura 4.24 muestra la salida gráfica de las nueve variables seleccionadas originalmente. Es importante notar la clara tendencia presente en algunas variables, mientras que otras pueden variar significativamente en sus escalas de representación.

Para garantizar una base sólida en este análisis es importante procesar los datos. En primer lugar, se ha eliminado la tendencia lineal de las variables seleccionadas para centrarnos en las variaciones subyacentes. Además, se ha estandarizado la escala de los datos para facilitar la comparación y el modelado.

Un paso importante en nuestra metodología es la división de los datos en dos conjuntos: el 70 % se reserva para el proceso de identificación de modelos, mientras que el resto se destina a la fase de validación. Esta división nos permite desarrollar y ajustar nuestros modelos utilizando una parte de los datos y luego evaluar su desempeño utilizando los datos de validación restantes. Esta estrategia es fundamental

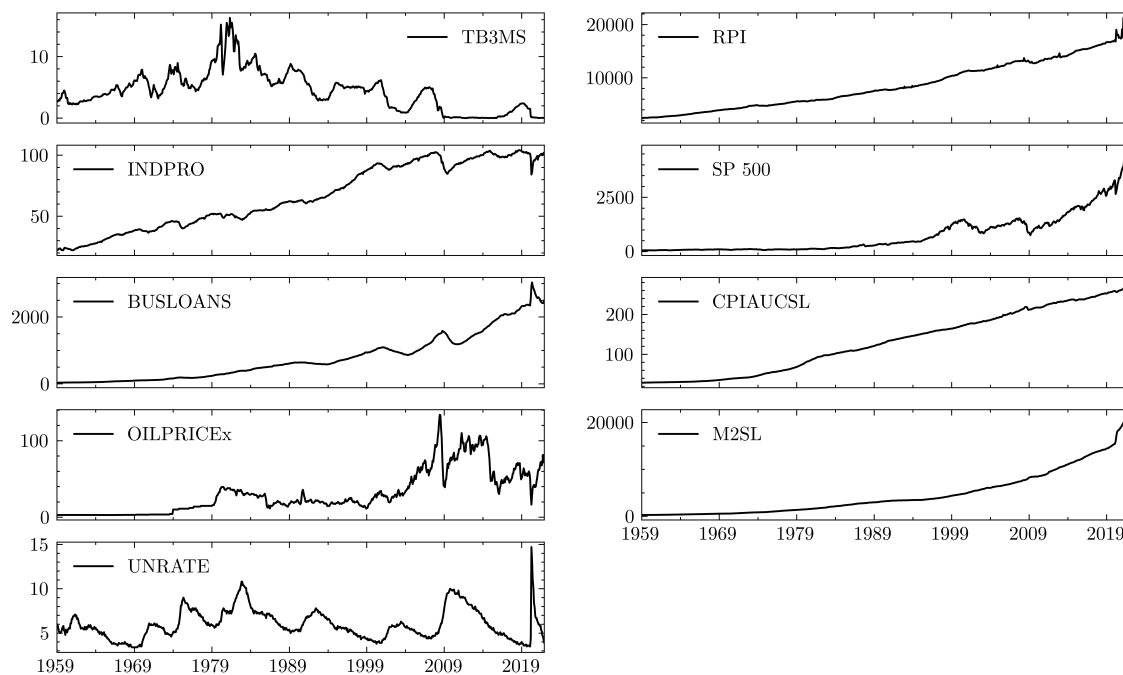


Figura 4.24: Salida gráfica de nueve variables macroeconómicas seleccionadas originalmente.

para asegurarnos de que nuestros modelos sean capaces de generalizar de manera efectiva a datos no vistos.

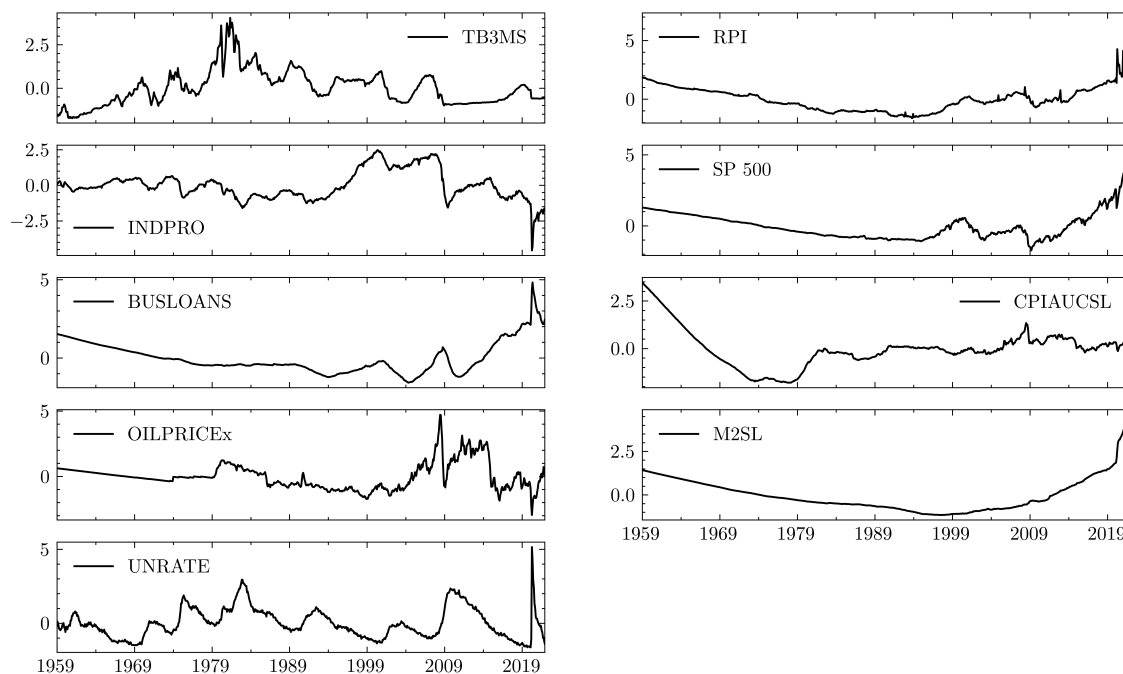


Figura 4.25: Salida gráfica de nueve variables macroeconómicas con tendencia lineal removida y estandarizadas.

La selección del orden del modelo es un paso crítico en el análisis de sistemas

dinámicos y series temporales. Como se explicó en la sección 2.5, nuestro enfoque consiste en analizar de forma gráfica los errores(desajuste) o residuos de los modelos calculados para diferentes valores de L . Si bien es cierto no existe un enfoque directo para seleccionar el orden, luego de hacer algunas consideraciones practicas, se decide utilizar un orden de $L = 2$.

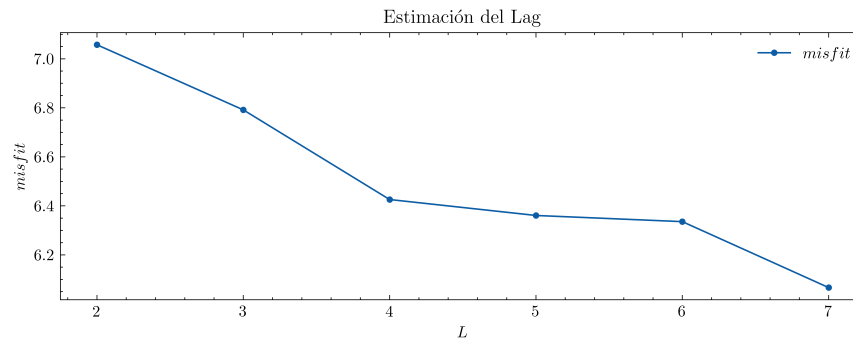


Figura 4.26: Desajuste (*misfit*) del modelo para diferentes valores de L . La gráfica nos indica que aumentar el orden del modelo no mejora la diferencia entre los errores en los conjuntos de identificación y validación.

Modelos identificados y análisis de resultados

1. Modelo de espacio de estados con SpLra.

$$A = \begin{bmatrix} 0 & 0.20 \\ 1 & 0.77 \end{bmatrix},$$

$$B = \begin{bmatrix} -0.02 & 0.09 & 0.29 & -0.04 & 0.37 & -0.33 & 0.01 & -0.02 \\ 0.04 & 0.00 & -0.35 & 0.11 & -0.45 & 0.33 & -0.01 & -0.06 \end{bmatrix},$$

$$C = \begin{bmatrix} 0 & 1.00 \end{bmatrix},$$

$$D = \begin{bmatrix} -0.08 & 0.00 & -0.56 & 0.00 & -0.08 & -0.07 & 0.06 & -0.08 \end{bmatrix}$$

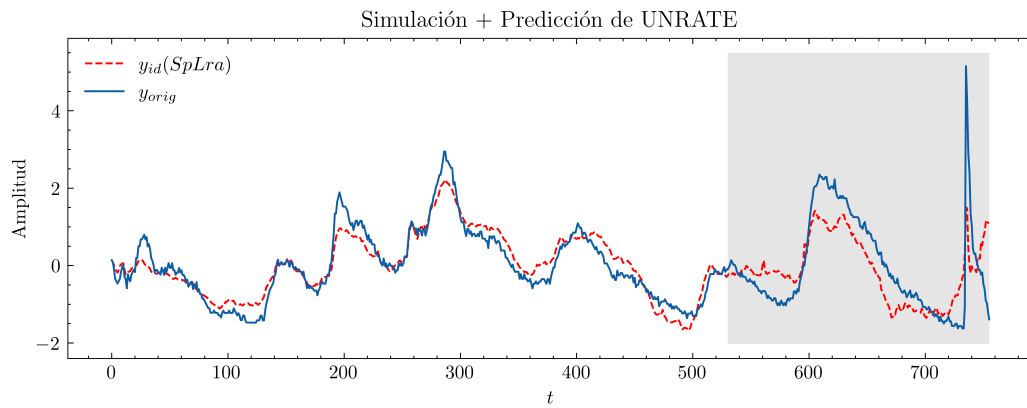


Figura 4.27: Salida gráfica del modelo identificado con SpLra.

2. Modelo de espacio de estados con `n4sid` de SIPPY.

$$A = \begin{bmatrix} 1.08 & -0.13 \\ 0.15 & 0.87 \end{bmatrix},$$

$$B = \begin{bmatrix} 0.002 & -0.001 & 0.01 & -0.03 & -0.03 & 0.04 & -0.01 & 0.18 \\ -0.001 & 0.07 & -0.001 & -0.06 & -0.03 & 0.03 & 0.03 & 0.20 \end{bmatrix},$$

$$C = \begin{bmatrix} 0.04 & -0.37 \end{bmatrix},$$

$$D = \begin{bmatrix} 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 \end{bmatrix}$$

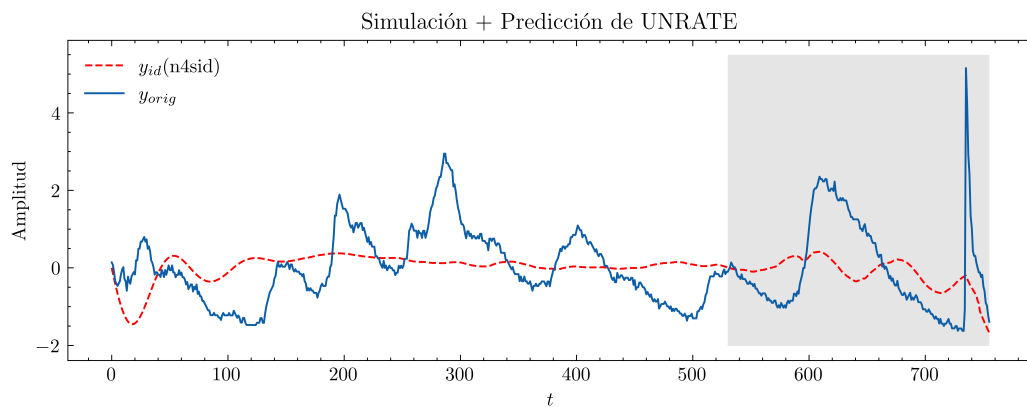


Figura 4.28: Salida gráfica del modelo identificado con `n4sid`.

3. Modelo AR(2) con variables *exógenas* con `AutoReg` del paquete `Statsmodels`

de Python. El modelo es de la forma

$$\hat{y}_{t+1} = \sum_{i=0}^1 \theta_{0,i} y_{t-i} + \theta_1 U_t$$

con

$$\theta_0 = \begin{bmatrix} 0.9336 & -0.0024 \end{bmatrix}^T$$

y

$$\theta_1 = \begin{bmatrix} 0.01 & 0.28 & -0.18 & 0.04 & -0.05 & -0.006 & 0.080 & -0.37 \end{bmatrix}^T$$

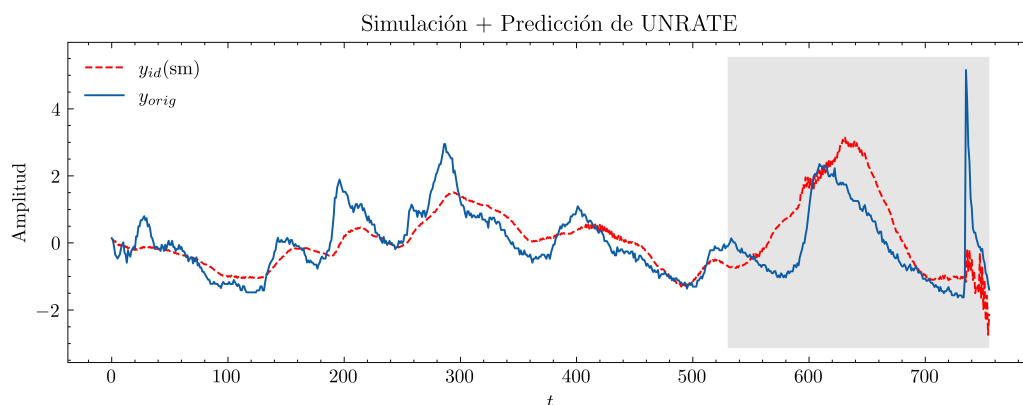


Figura 4.29: Salida gráfica del modelo identificado con AutoReg.

	Modelo					
	SpLra		N4sid		AutoReg	
nnz(%)	83		73		100	
métrica	id	val	id	val	id	val
%fit	88.26	23.00	0.00	4.00	45.00	4.00
rmse	0.31	0.95	0.92	1.18	0.49	1.17
R^2	0.88	0.40	0.00	0.07	0.69	0.084

Tabla 4.7: Resumen de resultados modelo econométrico base de datos FRED-MD.

Modelo de Orden Reducido (Factorial Dinámico)

Luego del pre-procesamiento de la base de datos FRED-MD, esta se reduce a un total de 105 series temporales con 756 entradas. Considerando los esquemas y algoritmos presentados en las Secciones 3.4,3.5, a continuación se construye un modelo reducido con $r = 9$ (factores latentes). La decisión de tomar nueve factores

se basa en los modelos anteriores donde se seleccionaron nueve variables exógenas. Los detalles y resultados de este experimento se pueden revisar y replicar utilizando el archivo `DFM.ipynb` (Jupyter Notebook) de [12].

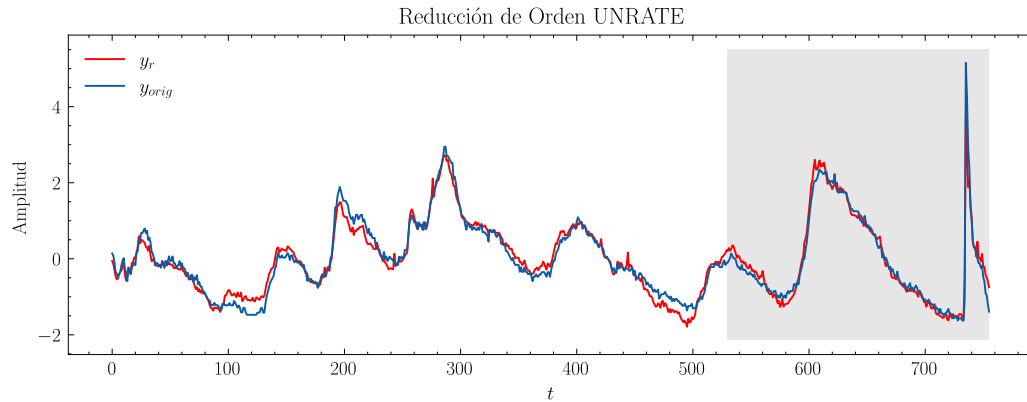


Figura 4.30: Salida gráfica de la serie UNRATE original (azul) y la serie generada a partir de reducción de orden con $r = 9$ factores latentes (rojo). La parte sombreada corresponde a los datos seleccionados para la fase de validación.

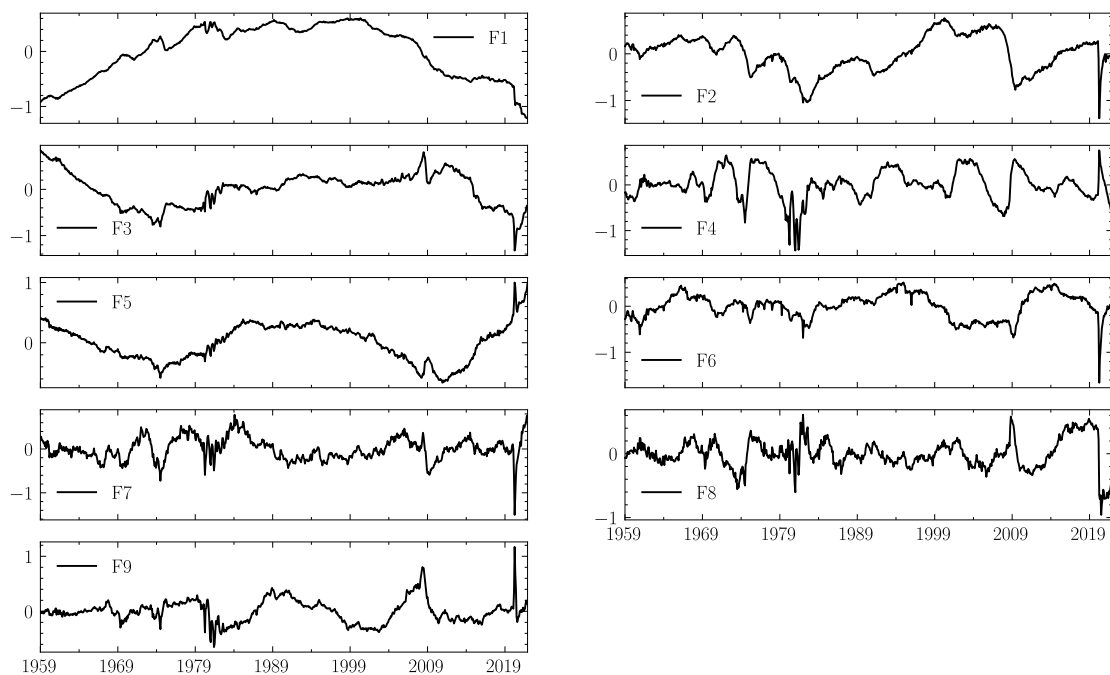


Figura 4.31: Salida gráfica de los factores $F \in \mathbb{R}^{9 \times 756}$ calculados a partir del Algoritmo 4.

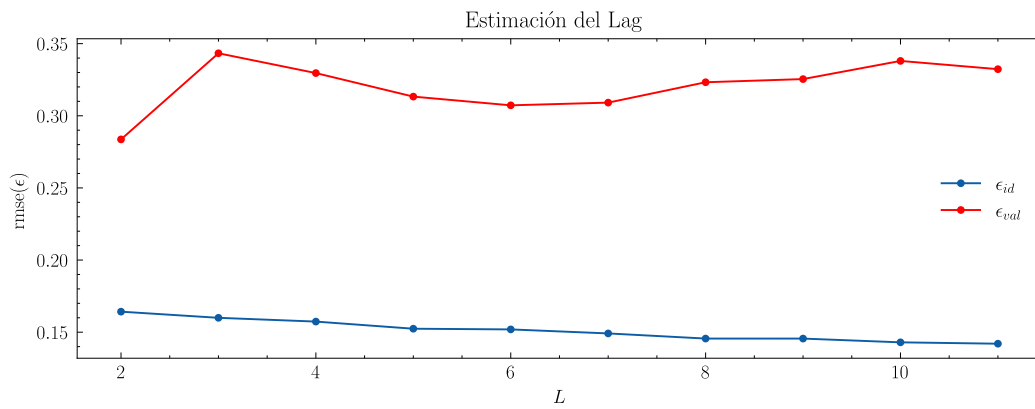


Figura 4.32: El gráfico muestra el $rmse$ en el conjunto de identificación (azul) y validación (rojo), para diferentes valores de L . Siguiendo con el método heurístico de la selección del orden del modelo, se toma $L = 6$.

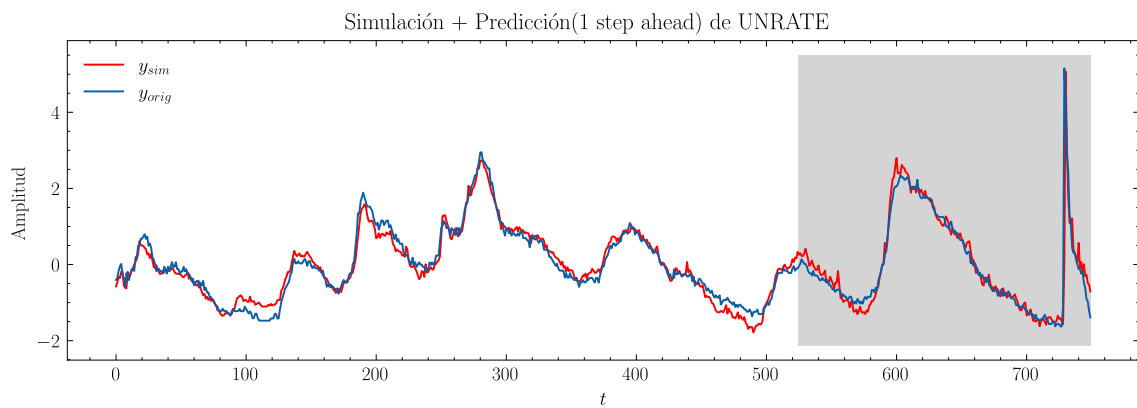


Figura 4.33: Salida gráfica de la serie UNRATE original (azul) y la serie generada a partir de reducción de orden con $r = 9$ factores latentes (rojo). La parte sombreada corresponde a los datos seleccionados para la fase de validación.

Capítulo 5

Discusión y Trabajo Futuro

5.1. Conclusiones

1. Los experimentos llevados a cabo, tanto en datos simulados como en datos reales, revelan resultados altamente prometedores del método propuesto **SPLra**. La integración de técnicas de esparcimiento, regularización y aproximación de bajo rango ha demostrado ser efectiva en múltiples aspectos del modelado de sistemas.
2. Una ventaja distintiva del método **SPLra**, en comparación con las alternativas evaluadas en este estudio, radica en su independencia del método de representación del sistema. Esto significa que **SPLra** se adapta sin problemas a una variedad de escenarios, ya sean sistemas **SISO** o **MIMO**, modelos expresados en variables de espacio de estados o función de transferencia. Esta flexibilidad lo convierte en una herramienta versátil y adaptable en el ámbito del modelado de sistemas.
3. Los resultados también evidencian que **SPLra** logra una aproximación precisa de las propiedades del sistema original cuando se aplican datos simulados. En el caso de datos reales, el método destaca por su capacidad para identificar modelos estables, lo que es crucial en aplicaciones del mundo real. Estos logros

- respaldan la utilidad y la efectividad del enfoque **SPLra** en la identificación y el modelado de sistemas complejos.
4. En resumen, los hallazgos de este estudio subrayan el potencial y las ventajas significativas del método **SPLra** en la selección de órdenes de modelo, el ajuste de datos y la aproximación de parámetros. Su versatilidad y su capacidad para mantener una representación precisa independientemente del tipo de sistema hacen que **SPLra** sea una herramienta valiosa en la investigación y la aplicación de modelos de sistemas dinámicos.

5.2. Trabajo Futuro

Como resultado de la presente investigación, es posible tomar diversas direcciones y oportunidades de investigación, que pueden extender y enriquecer aún más las técnicas desarrolladas en este trabajo. Una extensión natural y prometedora consiste en ampliar la clase de modelos considerados, explorando específicamente los *Polinomios Invariantes en el Tiempo*. Esta nueva clase de modelos nos permiten estudiar comportamientos no lineales y la posibilidad de analizar sistemas más complejos.

En el campo del control de sistemas puede proporcionar herramientas más avanzadas para el diseño de controladores robustos y adaptativos. Esto es especialmente relevante en sistemas complejos que requieren un control preciso y confiable.

Una extensión natural de esta investigación es la exploración y el desarrollo de modelos que tengan en cuenta específicamente los errores en las variables. Estos modelos, conocidos como modelos de EIV (por sus siglas en inglés), pueden ayudar a capturar y cuantificar de manera más precisa la incertidumbre asociada con las mediciones de entrada y salida en sistemas complejos.

La investigación futura puede enfocarse en el desarrollo de métodos de estimación robusta que sean capaces de lidiar de manera efectiva con la presencia de errores en las variables.

Bibliografía

- [1] (2000). 18 - the z-transform. In: A. Jeffrey, D. Zwillinger, eds., *Table of Integrals, Series, and Products (Sixth Edition)*. San Diego: Academic Press, sixth edition ed., pp. 1127–1132. doi:<https://doi.org/10.1016/B978-012294757-5/50022-2>.
- [2] Armenise, G., Vaccari, M., Di Capaci, R. B., Pannocchia, G. (2018). An open-source system identification package for multivariable processes. In: *2018 UKACC 12th International Conference on Control (CONTROL)*. pp. 152–157. doi:[10.1109/CONTROL.2018.8516791](https://doi.org/10.1109/CONTROL.2018.8516791).
- [3] Bai, J., Wang, P. (2016). Econometric Analysis of Large Factor Models. *Annual Review of Economics*, 8(1): 53–80. doi:[10.1146/annurev-economics-080315-015356](https://doi.org/10.1146/annurev-economics-080315-015356).
- [4] Berdiev, A. N., Saunoris, J. W. (2016). Financial development and the shadow economy: A panel VAR analysis. *Economic Modelling*, 57: 197–207. doi:[10.1016/j.econmod.2016.03.028](https://doi.org/10.1016/j.econmod.2016.03.028).
- [5] Bernanke, B. S., Boivin, J., Elias, P. (2005). Measuring the Effects of Monetary Policy: A Factor-Augmented Vector Autoregressive (FAVAR) Approach. *The Quarterly Journal of Economics*, 120(1): 387–422. doi:[10.1162/0033553053327452](https://doi.org/10.1162/0033553053327452).
- [6] Boutsidis, C., Magdon-Ismail, M. (2014). A note on sparse least-squares regression. *Information Processing Letters*, 114(5): 273–276. doi:[10.1016/j.ipl.2013.11.011](https://doi.org/10.1016/j.ipl.2013.11.011).

- [7] Boyd, S. P., Vandenberghe, L. (2018). *Introduction to applied linear algebra: vectors, matrices, and least squares*. Cambridge, UK ; New York, NY: Cambridge University Press.
- [8] Brumm, J., Scheidegger, S. (2017). Using Adaptive Sparse Grids to Solve High-Dimensional Dynamic Models. *Econometrica*, 85(5): 1575–1612. doi:10.3982/ECTA12216.
- [9] Buckmann, M., Joseph, A., Robertson, H. (2021). *Opening the Black Box: Machine Learning Interpretability and Inference Tools with an Application to Economic Forecasting*. Cham: Springer International Publishing, pp. 43–63. doi:10.1007/978-3-030-66891-4_3.
- [10] Cai, T. T., Zhang, C.-H., Zhou, H. H. (2010). Optimal rates of convergence for covariance matrix estimation. *The Annals of Statistics*, 38(4). doi:10.1214/09-AOS752.
- [11] Cao, J., Wang, J. (2019). Stock price forecasting model based on modified convolution neural network and financial time series analysis. *International Journal of Communication Systems*, 32(12): e3987. doi:10.1002/dac.3987.
- [12] Cardona, K. (2022). Sparse low rank approximation. <https://github.com/kerin84/SpLra>.
- [13] DeMoor, B. (1989). Daisy: Database for the identification of systems. Department of Electrical Engineering, ESAT/STADIUS, KU Leuven, Belgium.
- [14] Fan, J., Kim, D. (2017). Structured Volatility Matrix Estimation for Non-Synchronized High-Frequency Financial Data. *SSRN Electronic Journal*. doi:10.2139/ssrn.3085737.
- [15] Golub, G. H., Van Loan, C. F. (2013). *Matrix computations*. Johns Hopkins studies in the mathematical sciences. Baltimore: The Johns Hopkins University Press, fourth edition ed. OCLC: ocn824733531.

- [16] Heinrich, T. (2008). Identification problems in growth econometrics. *Stockholm University*.
- [17] Johansen, S. (2010). Some identification problems in the cointegrated vector autoregressive model. *Journal of Econometrics*, 158(2): 262–273. doi:10.1016/j.jeconom.2010.01.007.
- [18] Kaiser, E., Kutz, J. N., Brunton, S. L. (2018). Sparse identification of nonlinear dynamics for model predictive control in the low-data limit. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 474(2219): 20180335. doi:10.1098/rspa.2018.0335.
- [19] Kaufmann, S., Schumacher, C. (2019). Bayesian estimation of sparse dynamic factor models with order-independent and ex-post mode identification. *Journal of Econometrics*, 210(1): 116–134. doi:10.1016/j.jeconom.2018.11.008.
- [20] Kock, A. B., Tang, H. (2019). UNIFORM INFERENCE IN HIGH-DIMENSIONAL DYNAMIC PANEL DATA MODELS WITH APPROXIMATELY SPARSE FIXED EFFECTS. *Econometric Theory*, 35(2): 295–359. doi:10.1017/S0266466618000087.
- [21] Koopmans, T. C., Reiersol, O. (1950). The Identification of Structural Characteristics. *The Annals of Mathematical Statistics*, 21(2): 165–181. doi:10.1214/aoms/1177729837.
- [22] Mann, H. B., Wald, A. (1943). On Stochastic Limit and Order Relationships. *The Annals of Mathematical Statistics*, 14(3): 217–226. doi:10.1214/aoms/1177731415.
- [23] Markovsky, I., ed. (2006). *Exact and approximate modeling of linear systems: a behavioral approach*. Mathematical modeling and computation. Philadelphia: Society for Industrial and Applied Mathematics. OCLC: ocm62525241.
- [24] Markovsky, I. (2019). *Low-Rank Approximation: Algorithms, Implementation, Applications*. Communications and Control Engineering. Cham: Springer In-

- ternational Publishing. doi:10.1007/978-3-319-89620-5.
- [25] McCracken, M. W., Ng, S. (2016). FRED-MD: A Monthly Database for Macroeconomic Research. *Journal of Business & Economic Statistics*, 34(4): 574–589. doi:10.1080/07350015.2015.1086655.
- [26] Pierson, L. (2016). Python for data science.
- [27] Reddy, D. S., Prasad, P. R. C. (2018). Prediction of vegetation dynamics using NDVI time series data and LSTM. *Modeling Earth Systems and Environment*, 4(1): 409–419. doi:10.1007/s40808-018-0431-3.
- [28] Schaeffer, H., Tran, G., Ward, R. (2018). Extracting Sparse High-Dimensional Dynamics from Limited Data. *SIAM Journal on Applied Mathematics*, 78(6): 3279–3295. doi:10.1137/18M116798X.
- [29] Sims, C. A. (1980). Macroeconomics and Reality. *Econometrica*, 48(1): 1. doi:10.2307/1912017.
- [30] Soderstrom, T., Stoica, P. (1989). *System identification*. Prentice Hall International series in systems and control engineering. New York: Prentice Hall.
- [31] Stock, J. H., Watson, M. W. (2002). Macroeconomic Forecasting Using Diffusion Indexes. *Journal of Business & Economic Statistics*, 20(2): 147–162. doi:10.1198/073500102317351921.
- [32] Vides, F. (2019). On Cyclic Finite-State Approximation of Data-Driven Systems. In: *2019 IEEE 39th Central America and Panama Convention (CONCAPAN XXXIX)*. Guatemala City, Guatemala: IEEE, pp. 1–6. doi:10.1109/CONCAPANXXXIX47272.2019.8977039.
- [33] Vides, F. (2021). Sparse system identification by low-rank approximation. <https://arxiv.org/abs/2105.07522>.
- [34] Vides, F. (2021). A toolset with matlab, python and julia programs for approximate sparse system identification,. <https://github.com/FredyVides/SDSI>.

- [35] Vides, F. (2022). Computing semilinear sparse models for approximately eventually periodic signals. *IFAC-PapersOnLine*, 55(20): 205–210. doi:<https://doi.org/10.1016/j.ifacol.2022.09.096>. 10th Vienna International Conference on Mathematical Modelling MATHMOD 2022.
- [36] Willems, J. (1987). From time series to linear system—part iii: Approximate modelling. *Automatica*, 23(1): 87–115. doi:[10.1016/0005-1098\(87\)90120-8](https://doi.org/10.1016/0005-1098(87)90120-8).
- [37] Willems, J. C. (1986). From time series to linear system—part i. finite dimensional linear time invariant systems. *Automatica*, 22(5): 561–580. doi:[https://doi.org/10.1016/0005-1098\(86\)90066-X](https://doi.org/10.1016/0005-1098(86)90066-X).
- [38] Willems, J. C. (1986). From time series to linear system—part ii. exact modelling. *Automatica*, 22(6): 675–694. doi:[https://doi.org/10.1016/0005-1098\(86\)90005-1](https://doi.org/10.1016/0005-1098(86)90005-1).
- [39] Wold, H. (1938). On the inversion of moving averages. *Scandinavian Actuarial Journal*, 1938(3-4): 208–217. doi:[10.1080/03461238.1938.10413180](https://doi.org/10.1080/03461238.1938.10413180).
- [40] Xu, W., Peng, H., Zeng, X., Zhou, F., Tian, X., Peng, X. (2019). A hybrid modelling method for time series forecasting based on a linear regression model and deep learning. *Applied Intelligence*, 49(8): 3002–3015. doi:[10.1007/s10489-019-01426-3](https://doi.org/10.1007/s10489-019-01426-3).
- [41] Yule, G. U. (1927). VII. On a method of investigating periodicities disturbed series, with special reference to Wolfer’s sunspot numbers. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 226(636-646): 267–298. doi:[10.1098/rsta.1927.0007](https://doi.org/10.1098/rsta.1927.0007).
- [42] Zhang, G. (2003). Time series forecasting using a hybrid ARIMA and neural network model. *Neurocomputing*, 50: 159–175. doi:[10.1016/S0925-2312\(01\)00702-0](https://doi.org/10.1016/S0925-2312(01)00702-0).