

2020

IDENTIFICACION DE SISTEMAS DINAMICOS LINEALES MEDIANTE MAXIMA VEROSIMILITUD CON MEZCLA FINITA DE DISTRIBUCIONES NORMAL

BITTNER HOFFMANN, GUSTAVO JUAN ESTEBAN

<https://hdl.handle.net/11673/49513>

Repositorio Digital USM, UNIVERSIDAD TECNICA FEDERICO SANTA MARIA

UNIVERSIDAD TÉCNICA FEDERICO SANTA MARÍA

DEPARTAMENTO DE ELECTRÓNICA
VALPARAÍSO, CHILE



UNIVERSIDAD TECNICA
FEDERICO SANTA MARIA

**IDENTIFICACIÓN DE SISTEMAS DINÁMICOS
LINEALES MEDIANTE MÁXIMA VEROSIMILITUD
CON MEZCLA FINITA DE DISTRIBUCIONES
NORMALES**

Tesis de Grado presentada por

Gustavo Juan Esteban Bittner Hofmann

como requisito parcial para optar al título de

Ingeniero Civil Electrónico

y al grado de

Magíster en Ciencias de la Ingeniería Electrónica

Profesor Guía

Dr. Juan Carlos Agüero Vásquez

Profesor Co-guía

Dr. Rodrigo Carvajal Guerra

Julio 2020

UNIVERSIDAD TÉCNICA FEDERICO SANTA MARÍA

Departamento de Electrónica

**Identificación de sistemas dinámicos lineales
mediante Máxima Verosimilitud con mezcla
finita de distribuciones Normales**

Tesis presentada por

Gustavo Juan Esteban Bittner Hofmann

como requisito parcial para optar al título de

Ingeniero Civil Electrónico

y al grado de

Magíster en Ciencias de la Ingeniería Electrónica

Profesor Guía

Dr. Juan Carlos Agüero Vásquez.

Profesor Co-guía

Dr. Rodrigo Carvajal Guerra.

Valparaíso, Julio 2020.

Título de Tesis:

**IDENTIFICACIÓN DE SISTEMAS DINÁMICOS LINEALES MEDIAN-
TE MÁXIMA VEROSIMILITUD CON MEZCLA FINITA DE DIS-
TRIBUCIONES NORMALES**

Autor:

Gustavo Juan Esteban Bittner Hofmann

TRABAJO DE TESIS, presentada en cumplimiento parcial como requisito para optar al título de Ingeniero Civil Electrónico y al grado de Magíster en Ciencias de la Ingeniería Electrónica de la Universidad Técnica Federico Santa María.

Profesor Guía,

Dr. Juan Carlos Agüero Vásquez.

Profesor Co-Guía,

Dr. Rodrigo Carvajal Guerra.

Examinador Interno,

Dr. Francisco Vargas Parra.

Examinador Externo,

Dr. Boris Godoy Torres.

Valparaíso, Julio 2020.

*A mi Mamá, Papá y Hermana,
por su sacrificio, comprensión,
dedicación, apoyo y amor.*

AGRADECIMIENTOS

Este trabajo es el resultado de un largo viaje, algo complicado, con pausas, sorpresas, alegrías y algunos momentos oscuros que me han permitido crecer y comprender que irónicamente es necesaria la adversidad para apreciar lo realmente importante.

Sin embargo, nunca he estado solo, por el contrario, he contado con numerosas personas que han aportado en mayor o menor medida, cada cual a su manera, forjando quien soy y en consecuencia siendo parte fundamental de esta historia.

En primer lugar, debo dar gracias a mi familia: mis padres, mi hermana y mis pequeños sobrinos por el apoyo, guía, cariño, enseñanzas, alegrías y confianza brindadas, además de enseñarme que el esfuerzo y el sacrificio son esenciales para lograr mis metas. Quiero agradecer además a mi Oma Inés, a mi tía Marlys y su familia, quienes a pesar de la distancia, han sido parte fundamental de todo este proceso. Si bien para muchos puede ser obvio, pero la paciencia, el cariño, la constancia y la infinita entrega que me han brindado, construyeron gran parte de la persona en la cual me he convertido. Sin ellos nada de esto sería posible.

También debo agradecer a mis otras familias. A aquellos que me acogieron en sus casas, entregándome todo aquello que la distancia me había privado. Mi tía Sonia Ríos, Carmen Aguilar y Mario Torres, quienes en mi infancia, me trataron como a uno más de su familia, brindándome cariño y cuidado. También a mi tía Elena Yopez y Jaime Morales, quienes me acogieron cuando debí alejarme nuevamente y llegué a Valparaíso a estudiar. Ambas familias son mi familia. Los siento como parte importante de todo lo que soy ahora y por eso les estoy eternamente agradecido.

En lo académico, también he descubierto personas excepcionales. Mi profundo agradecimiento a los profesores Juan Carlos Agüero y Rodrigo Carvajal, quienes me introdujeron en el mundo de la investigación científica, guiándome cuando sentía que estaba desorientado. Agradezco su apoyo y motivación, alentándome en este camino, que ya había recorrido alguna vez bajo la guía de Ricardo Rojas, Cristian Rojas y Patricio Valenzuela. Al igual que Rafael Orellana, todos ellos me han enseñado y apoyado, contribuyendo en gran parte al desarrollo de este trabajo.

Tampoco puedo dejar de mencionar a quienes, ingresando en mi vida como profesos-

res, se convirtieron en más que eso. Ricardo Rojas, Mario Salgado (q.e.p.d.) y Eduardo Silva (q.e.p.d.) quienes no solo me entregaron herramientas en lo académico, sino que me enseñaron, aconsejaron y compartieron su sabiduría más allá de las aulas, orientándome en mi vida. Muchísimas gracias por ser parte de mi crecimiento intelectual y personal. Espero estén orgullosos y felices del resultado. Jamás podré olvidarlos.

Gracias, a Elena Garcés, Soledad Tilleria (q.e.p.d.), José Aros, Juan Carlos Gutiérrez, quienes en el colegio me entregaron la base de todo lo que hoy he construido, despertando mi curiosidad y hambre de conocimiento a través de preguntas, alentandome a encontrar respuestas. Gracias también a Juan Yuz, Jaime Glaría, Cesar Silva, Juan Hernandez, Leopoldo Silva (q.e.p.d.), Sergio Olavarria, Samir Kouro, Daniel Rodríguez, Victor Gonzalez, Wilmer Bazan, Milan Derpich, quienes expandieron los límites de mi curiosidad, alimentando mis ansias de aprendizaje y descubrimiento.

Más allá de lo académico, agradezco también a quienes han sido parte fundamental de este trabajo. Felipe Santander, Paola Antoine, Gabriel Urrutia, Nicolas Müller, Ignacio Ferruzola, Osvaldo Olguin, Tania Rodriguez, Juan Zepeda, mis amigos, esos que me han visto en lo más hondo, que me tendieron la mano en numerosas oportunidades. Me sacaron de lo más profundo de la oscuridad, mostrándome cariño y lealtad incondicionales. También a quien encontré sin buscar, a Veronica Cortés-Monroy quien me ha acompañado, apoyado y aconsejado. Sé que están ahí, pase lo que pase. A ellos solo agradecer la paciencia, el tiempo, las alegrías, los consejos, el apoyo y la entrega. Espero sepan que pueden contar conmigo para lo que sea.

A Iván Toro y Eugenio Urzúa, quienes eran mis compañeros en la pensión, formando parte de mi infancia con los más bellos recuerdos de una época lejana.

No puedo dejar de mencionar a Matías García, Javiera Quintana, Jesu Olavarria, Dany Sateler, Sebastián Araya, Gabriel Galindo, Matias Müller, Alejandro Krebs, Felipe Maira, Nelly Boldt, Constanza Manriquez, Susan Sulzer. Gracias a cada uno de ustedes por los consejos, las enseñanzas, el apoyo y los buenos momentos compartidos.

Una mención especial a Fabiola Narváez, Fabiola Jaramillo y Jorge Larraguibel. Gracias a su trabajo, dedicación y paciencia he crecido personalmente y me encuentro mejor preparado para las batallas venideras.

Ahora escribiendo me doy cuenta que son tantas las personas a las que debo nombrar,

Rodrigo Saez, Maira Fritz, Marthin Neumann, Thomas Neumann, Constanza Yunge, Francisco Guentelican, Eduardo Toro, Allyson Gehrkie, Erika Stange, Maura Ampuero, Egon Werner, Claudio Bittner, Alicia Saavedra y María Muñoz (q.e.p.d.). A todos ustedes, gracias por la oportunidad de ser parte de sus vidas.

Finalmente quisiera agradecer a la Dirección General de Investigación y Postgrado (DGIP) y a la Comisión Nacional de Ciencia y Tecnología (CONICYT), en particular a CONICYT-PCHA/ Beca Magíster Nacional/Año 2014 - Folio 2140824, por el apoyo económico entregado durante este proceso.

Además, este trabajo fue apoyado parcialmente por FONDECYT a través del proyecto No 1181158, el programa PIIC 05/2019 de la Universidad Técnica Federico Santa María y el Centro Avanzado de Ingeniería Eléctrica y Electrónica (AC3E, Proyecto Basal FB0008), Chile.

«La verdadera amistad es planta de lento crecimiento que debe sufrir y vencer los embates del infortunio antes de que sus frutos lleguen a completa madurez»

— George Washington

Maximum Likelihood identification for Linear Dynamic Systems with finite Gaussian mixture noise distribution

Thesis submitted by

Gustavo Juan Esteban Bittner Hofmann

in partial fulfillment of the requirements for the professional title of

Electronic Engineer

and the degree of

Master of Science in Electronic Engineering

Thesis supervisor

Dr. Juan Carlos Agüero Vásquez

Thesis co-supervisor

Dr. Rodrigo Carvajal Guerra

Internal Reviewer

Dr. Francisco Vargas Parra

External Reviewer

Dr. Boris Godoy Torres

ABSTRACT

This Thesis addresses the identification problem of output-error systems with non-Gaussian measurement noise.

Initially the problem is restricted to minimum phase noise zeros, i.e. the roots of the numerator associated with the noise transfer function lies within the unit circle.

The noise distribution is approximated by a finite Gaussian mixture, whilst the parameters of the system and the parameters that approximate the noise distribution are simultaneously estimated using the principle of Maximum Likelihood.

To this end, a global optimization algorithm is utilized to solve the resulting non-convex optimization problem.

It is shown that our approach improves the accuracy of the estimates, when compared with classic estimation techniques such as the Prediction Error Method (PEM), in

terms of covariance of the estimation error, while also obtaining an approximation of the noise distribution. The benefits of the proposed technique are illustrated by numerical simulations.

Later, a Maximum Likelihood estimation algorithm for a non-minimum-phase linear dynamic system with Gaussian mixture noise distribution is developed.

Based on the Expectation-Maximization algorithm, we propose an identification technique to estimate the system model parameters and the Gaussian mixture parameters. We show that the estimates obtained by using this approach exhibit good accuracy. The benefits of our proposal are illustrated via numerical simulations.

The work hereby presented is divided in three main parts. First, an overview of classical system identification and the state of the art technique, given by the Method of Moments, for solving the problem of interest is discussed. The second part of the thesis presents a solution based on a global optimization method, given by the Pattern Search algorithm, to solve the Maximum Likelihood estimation problem. A Gaussian Mixture Model is considered to approximate the noise distribution and its parameters are simultaneously estimated with the system parameters. The third part of the thesis addresses an extension by considering non-minimum phase noise zeros. An Expectation-Maximization based algorithm is proposed to estimate the system model parameters and the Gaussian mixture parameters.

Keywords – Gaussian Mixture Model, Linear Dynamical Systems, Maximum Likelihood, non-Gaussian Noise Distribution, Minimum Phase Zeros, non-Minimum Phase Zeros, Expectation-Maximization, Method of Moments, Global Optimization.

ÍNDICE

	Página
ÍNDICE	x
ÍNDICE DE FIGURAS	xi
ÍNDICE DE TABLAS	xiii
ÍNDICE DE ALGORITMOS	xv
1. INTRODUCCIÓN	1
1.1. Definición del problema	2
1.2. Estado del arte	3
1.3. Organización de la tesis	5
1.4. Publicaciones realizadas por el autor	7
1.5. Notación	9
1.5.1. Glosario	9
1.5.2. Abreviaciones	10
1.5.3. Convenciones de Notación	11
1.5.4. Convenciones de Polinomios	11
2. IDENTIFICACIÓN DE SISTEMAS	13
2.1. Sistema	13
2.2. Elementos principales de identificación de sistemas	14
2.2.1. Conjunto de Datos	14
2.2.2. Conjunto de Modelos	15
2.2.3. Método de Identificación	16
2.3. Método de máxima verosimilitud (ML)	17
2.3.1. Análisis asintótico	19

2.4. Método del error de predicción (PEM)	20
2.5. Algunas definiciones	24
2.6. Lecturas complementarias	28
3. MÉTODO DE LOS MOMENTOS	30
3.1. Descripción del problema	30
3.2. Supuestos del Modelo	32
3.3. Preliminares	33
3.3.1. Momentos y cumulantes	33
3.3.2. Cumulantes en procesos aleatorios y sistemas	36
3.4. Procedimiento de Identificación	39
3.4.1. Estimación de un modelo espectralmente equivalente	39
3.4.2. Determinación de la ubicación correcta de los ceros	41
3.5. Lecturas complementarias	43
4. IDENTIFICACIÓN MEDIANTE ML DE SISTEMAS CON RUIDO	
GMM	46
4.1. Estimación de una constante	46
4.1.1. Estimación mediante mínimos cuadrados (LS)	47
4.1.2. Estimación mediante máxima verosimilitud (ML)	48
4.1.3. Comparación de precisión entre ambos estimadores	49
4.2. Función de verosimilitud para ruido no-gaussiano	50
4.2.1. Modelo de mezcla de distribuciones gaussianas para aproximar ruido no-gaussiano	50
4.2.2. Función de verosimilitud para mezcla de distribuciones gaussianas	51
4.3. Extensión a Sistemas Dinámicos Lineales	53
4.4. Ejemplos numéricos	55
4.4.1. Configuración de las simulaciones	56
4.4.2. Criterios de desempeño	56
4.4.3. Ejemplo 1: Sistema dinámico con ruido de distribución uniforme	57
4.4.4. Ejemplo 2: Sistema dinámico con ruido de distribución χ^2	63
4.4.5. Resumen y análisis de los resultados	69
5. IDENTIFICACIÓN MEDIANTE ML DE SISTEMAS DE FASE NO-MÍNIMA CON RUIDO GMM	73
5.1. Descripción del problema	74
5.1.1. Supuestos	75

5.2. Estimación de sistemas de fase no-mínima con mezcla de distribuciones gaussianas mediante máxima verosimilitud	76
5.2.1. Error de predicción para sistemas dinámicos de fase no-mínima . .	77
5.2.2. Estimación de sistemas lineales de fase no-mínima utilizando un modelo de mezcla gaussiana mediante máxima verosimilitud	78
5.3. Algoritmo de maximización de la esperanza (EM) para sistemas lineales con mezcla finita de distribuciones gaussianas	80
5.3.1. Formulación del algoritmo basado en la maximización de la esperanza EM	80
5.3.2. Procedimiento de inicialización para el algoritmo EM con GMM .	85
5.4. Ejemplos numéricos	86
5.4.1. Configuración de las simulaciones	87
5.4.2. Ejemplo 1: Sistema de fase no-mínima con distribución de ruido dada por una mezcla gaussiana superpuesta	87
5.4.3. Ejemplo 2: Sistema de fase no-mínima con ruido de distribución uniforme	89
5.4.4. Ejemplo 3: Sistema de fase mínima con ruido de distribución Chi-cuadrado	91
6. CONCLUSIONES	94
6.1. Resumen de los capítulos de la tesis destacando las contribuciones	94
6.2. Conclusiones	95
6.3. Trabajos futuros	96
REFERENCIAS	97
A. APÉNDICE	115
A.1. Probabilidades y procesos aleatorios	115
A.2. Algoritmo maximización de la esperanza EM	116
A.2.1. Ejemplo numérico	120
A.2.2. Lecturas complementarias	121
B. ALGORITMOS DE OPTIMIZACIÓN GLOBAL	124
B.1. Breve Introducción	124
B.2. Recocido Simulado (Simulated Annealing)	127
B.2.1. Inspiración	127
B.2.2. Metáfora	128
B.2.3. Estrategia y procedimiento	128

B.2.4.	Algunas consideraciones	129
B.2.5.	Lecturas complementarias	130
B.3.	Algoritmo genético (Genetic Algorithm)	130
B.3.1.	Inspiración	131
B.3.2.	Metáfora	131
B.3.3.	Estrategia y procedimiento	131
B.3.4.	Algunas consideraciones	135
B.3.5.	Lecturas complementarias	136
B.4.	Optimización mediante enjambre de partículas (Particle Swarm Opti- zation)	136
B.4.1.	Inspiración	136
B.4.2.	Metáfora	137
B.4.3.	Estrategia y procedimiento	137
B.4.4.	Algunas consideraciones	139
B.4.5.	Lecturas complementarias	139
B.5.	Búsqueda de patrones (Pattern Search)	140
B.5.1.	Estrategia y procedimiento	140
B.5.2.	Algunas consideraciones	141
B.5.3.	Lecturas complementarias	142
B.6.	Comparación entre algoritmos	144
B.6.1.	Funciones de prueba	144
B.6.2.	Pruebas realizadas en $\mathbb{R}^2$	148
B.6.3.	Pruebas realizadas en $\mathbb{R}^n$; $n \in [2, 20]$	149
B.7.	Códigos utilizados	152
B.7.1.	Recocido simulado (Simulated Annealing)	152
B.7.2.	Algoritmo genético (Genetic Algorithm)	152
B.7.3.	Búsqueda de patrones (Pattern Search)	152
B.7.4.	Optimización mediante enjambre de partículas (Particle Swarm Optimization)	152

ÍNDICE DE FIGURAS

2.1. Diagrama de bloques de un sistema genérico.	14
3.1. Cumulantes muestrales $\hat{C}_k(x)$ versus cumulantes teóricos $C_k(x)$ a partir de n realizaciones.	36
3.2. Varianza teórica $P_{C_k(x)}$ y muestral $\hat{P}_{C_k(x)}$ de los cumulantes muestrales a partir de n realizaciones.	37
4.1. Covarianzas del error de estimación para el ejemplo de estimar una constante cuando la distribución del ruido corresponde a un GMM de dos componentes.	50
4.2. Logaritmo de la función de verosimilitud para valores medios μ_i y μ_j con $i \neq j$ al utilizar GMM para aproximar la distribución del ruido.	53
4.3. Logaritmo de la función de verosimilitud para el valor medio μ_j del GMM y el parámetro θ	54
4.4. Aproximaciones de la distribución uniforme del ruido al utilizar un GMM para el Ejemplo 1 con $N = 100$ datos.	58
4.5. Aproximaciones de la distribución uniforme del ruido al utilizar un GMM para el Ejemplo 1 con $N = 500$ datos.	59
4.6. Aproximaciones de la distribución uniforme del ruido al utilizar un GMM para el Ejemplo 1 con $N = 1000$ datos.	60
4.7. Aproximaciones de la distribución uniforme del ruido al utilizar un GMM para el Ejemplo 1 con $N = 5000$ datos.	61
4.8. Aproximaciones de la distribución uniforme del ruido al utilizar un GMM para el Ejemplo 1 con $N = 10000$ datos.	62
4.9. Aproximaciones de la distribución <i>Chi-cuadrado</i> del ruido al utilizar un GMM para el Ejemplo 2 con $N = 100$ datos.	64
4.10. Aproximaciones de la distribución <i>Chi-cuadrado</i> del ruido al utilizar un GMM para el Ejemplo 2 con $N = 500$ datos.	65

4.11. Aproximaciones de la distribución <i>Chi-cuadrado</i> del ruido al utilizar un GMM para el Ejemplo 2 con $N = 1000$ datos.	66
4.12. Aproximaciones de la distribución <i>Chi-cuadrado</i> del ruido al utilizar un GMM para el Ejemplo 2 con $N = 5000$ datos.	67
4.13. Aproximaciones de la distribución <i>Chi-cuadrado</i> del ruido al utilizar un GMM para el Ejemplo 2 con $N = 10000$ datos.	68
5.1. Diagrama de bloques del esquema utilizado para realizar el filtrado anti-causal y obtener $\varepsilon_t(\theta)$ en (5.10).	77
5.2. Histograma de $\varepsilon_t(\hat{\theta}^{(0)})$ obtenido a partir de (5.9) y (5.10), con $\hat{\theta}^{(0)} = \hat{\theta}^{(\text{PEM})}$	86
5.3. Resultados de la estimación paramétrica para el Ejemplo 1.	89
5.4. Aproximación de la distribución del ruido ω_t al utilizar un GMM para el Ejemplo 1.	89
5.5. Resultados de la estimación paramétrica para el Ejemplo 2.	90
5.6. Aproximación de la distribución del ruido ω_t al utilizar un GMM para el Ejemplo 2.	91
5.7. Resultados de la estimación paramétrica para el Ejemplo 3.	92
5.8. Aproximación de la distribución del ruido ω_t al utilizar un GMM para el Ejemplo 3.	92
A.1. Logaritmo de la función de verosimilitud $\ell_\theta(Y)$ y función auxiliar $\mathcal{Q}(\theta, \theta_k)$ para diferentes iteraciones k del algoritmo EM para el sistema (A.23) . . .	121
B.1. Óptimo global y local.	125
B.2. Función convexa.	126
B.3. Ejemplos de recombinación binaria.	133
B.4. Mutación binaria.	134
B.5. Comportamiento del algoritmo de optimización mediante búsqueda de patrones (PS) para diferentes iteraciones.	142
B.6. Funciones de prueba utilizadas en $\mathbb{R}^2$	147
B.7. Error medio $\bar{\mu}_e$ de las soluciones óptimas obtenidas utilizando diferentes algoritmos sobre la función de Rosenbrock en $\mathbb{R}^n$, $n \in [2, 20]$	150
B.8. Error medio $\bar{\mu}_e$ de las soluciones óptimas obtenidas utilizando diferentes algoritmos sobre la función de Rastrigin en $\mathbb{R}^n$, $n \in [2, 20]$	150
B.9. Error medio $\bar{\mu}_e$ de las soluciones óptimas obtenidas utilizando diferentes algoritmos sobre la función de Ackley en $\mathbb{R}^n$, $n \in [2, 20]$	151

ÍNDICE DE TABLAS

2.1. Algunas clasificaciones de los sistemas LTI de acuerdo a su estructura. . .	16
4.1. Estadísticas de la simulación de Monte-Carlo para el Ejemplo 1 con $N = 100$.	59
4.2. Estadísticas de la simulación de Monte-Carlo para el Ejemplo 1 con $N = 500$.	60
4.3. Estadísticas de la simulación de Monte-Carlo para el Ejemplo 1 con $N =$ 1000.	61
4.4. Estadísticas de la simulación de Monte-Carlo para el Ejemplo 1 con $N =$ 5000.	62
4.5. Estadísticas de la simulación de Monte-Carlo para el Ejemplo 1 con $N =$ 10000.	63
4.6. Estadísticas de la simulación de Monte-Carlo para el Ejemplo 2 con $N = 100$.	64
4.7. Estadísticas de la simulación de Monte-Carlo para el Ejemplo 2 con $N = 500$.	65
4.8. Estadísticas de la simulación de Monte-Carlo para el Ejemplo 2 con $N =$ 1000.	66
4.9. Estadísticas de la simulación de Monte-Carlo para el Ejemplo 2 con $N =$ 5000.	67
4.10. Estadísticas de la simulación de Monte-Carlo para el Ejemplo 2 con $N =$ 10000.	68
4.11. Resumen de las estadísticas de la simulación de MC para el Ejemplo 1 (ver página 57).	69
4.12. Resumen de las estadísticas de la simulación de MC para el Ejemplo 2 (ver página 63).	69
4.13. Tiempos medios de ejecución t_e [s] para los Ejemplos 1 y 2.	69
5.1. Estadísticas de la simulación de Monte-Carlo para el Ejemplo 1.	88
5.2. Estadísticas de la simulación de Monte-Carlo para el Ejemplo 2.	90
5.3. Estadísticas de la simulación de Monte-Carlo para el Ejemplo 3.	92

A.1. Estadísticas de la simulación de Monte-Carlo al estimar θ mediante el algoritmo EM para el sistema (A.23).	120
B.1. Tiempos de ejecución promedio $t_e[s]$.	148
B.2. Error promedio μ_e .	149
B.3. Desviación estándar del error σ_e .	149
B.4. Tiempos de ejecución promedio $t_e[s]$.	151

ÍNDICE DE ALGORITMOS

A.2.1Pseudo-código para el algoritmo maximización de la esperanza (EM). . . .	119
B.2.1Pseudo-código para el algoritmo Recocido Simulado (SA).	129
B.3.1Pseudo-código para el algoritmo genético (GA)	135
B.4.1Pseudo-código para el algoritmo de optimización mediante enjambre de partículas (PSO).	138
B.5.1Pseudo-código para el algoritmo de optimización mediante búsqueda de patrones (PS).	143

INTRODUCCIÓN

La **identificación de sistemas** representa una fase crucial en numerosas aplicaciones de la ingeniería. Se entiende por identificación de sistemas a la obtención de modelos matemáticos a partir de datos experimentales (típicamente entrada-salida), los cuales generalmente se encuentran contaminados con ruido [1–3].

Los algoritmos de identificación involucran generalmente sistemas con características particulares, por ejemplo embebido estocástico (*Stochastic Embedding*) [4], sistemas lineales [5], sistemas con salida cuantizada [6, 7], sistemas distribuidos [8], sistemas continuos [9] entre otros. Dichos enfoques poseen supuestos similares respecto de las características de las fuentes de ruido, por ejemplo media cero, función de densidad de probabilidad gaussiana, entre otras. No obstante, existen sistemas donde la distribución del ruido es **no-gaussiana**, en los cuales los algoritmos de identificación clásicos basados en el supuesto de distribución normal poseen un desempeño deficiente, particularmente cuando se utilizan pocos datos (ver e.g. [10, 11])

Los algoritmos clásicos de identificación, como **máxima verosimilitud** (ML), han sido ajustados considerando fuentes de ruido con distribuciones gaussianas, dando origen a los **métodos del error de predicción** (PEM) para sistemas lineales [1, 2]. Bajo este supuesto sobre la distribución del ruido, dichos estimadores son **asintóticamente consistentes**, es decir, que la estimación mejora a medida que el número de datos se incrementa, aproximándose hacia el valor verdadero de los parámetros del sistema θ_0 .

En la literatura se ha visto cómo la **eficiencia** de los estimadores disminuye cuando no se satisface la suposición de gaussianidad del ruido de medición. Por ejemplo, en [12] se muestra que un sistema autoregresivo (AR) excitado con ruido gaussiano presenta una **cota inferior de Cramér-Rao** (CRB) superior (peor) que para el caso no-gaussiano. En [13] se presenta una extensión de dicho resultado para sistemas autorregresivos de media móvil (ARMA). Estos resultados implican que es posible estimar los parámetros del sistema con mayor precisión en el caso no-gaussiano.

El objetivo principal de este trabajo es desarrollar un estimador para sistemas lineales, cuando la distribución del ruido es desconocida y no-gaussiana. A continuación se

define formalmente el problema de interés.

1.1. Definición del problema

Este trabajo considera la identificación de un sistema lineal (LTI) cuando la distribución del ruido es **no-gaussiana**. Considere entonces un sistema lineal genérico dado por

$$y_t = G(q^{-1}, \theta_0) u_t + H(q^{-1}, \theta_0) w_t, \quad (1.1)$$

donde y_t corresponde a la salida del sistema, u_t a la entrada determinística (manipulable) del sistema, G y H son funciones de transferencia estables del operador retardo de tiempo q^{-1} parametrizadas mediante θ , mientras que w_t es ruido blanco de distribución cualesquiera y desconocida.

Observación 1.1. *Una distribución es generalmente denominada **no-gaussiana**, debido a que no se utiliza una distribución particular y se asume desconocida. El problema de identificación del sistema dado por (1.1), cuando w_t posee una distribución arbitraria conocida ha sido resuelto con anterioridad, obteniéndose una versión particular de PEM o ML para la distribución en cuestión. [1, 2].* ∇

Utilizando el **teorema de aproximación de Wiener** [14, 15] es posible aproximar arbitrariamente la distribución desconocida del ruido w_t , como una combinación convexa (combinación lineal donde todos los coeficientes son no negativos y suman 1) de distribuciones Gaussianas, es decir

$$p(w_t) \approx \sum_{j=1}^M \alpha_j \phi(w_t; \mu_j, \Sigma_j) \quad \text{s.t.} \quad \sum_{j=1}^M \alpha_j = 1, \quad (1.2)$$

para algún $M \in \mathbb{N}$; $\phi(w_t; \mu_j, \Sigma_j)$ representa la distribución gaussiana correspondiente a la j -ésima componente de la mezcla, con media μ_j y covarianza Σ_j , mientras que α_j corresponde al j -ésimo coeficiente de peso de la mezcla.

En primera instancia se restringe el problema al considerar solo funciones de transferencia de **fase mínima** (todos los ceros se encuentran dentro del círculo unitario) en las dinámicas del ruido, es decir $H(q^{-1}, \theta_0)$. Utilizando el método de **máxima verosimilitud** se busca estimar el vector de parámetros aumentado $\beta = \{\theta, \eta\}$, donde θ corresponden a los parámetros del sistema y $\eta = \{\alpha_j, \mu_j, \Sigma_j\}_{j=1}^M$ los parámetros que aproximan la distribución del ruido como una mezcla de distribuciones gaussianas (ver (1.2)).

Puesto que el problema de optimización resultante, al utilizar el estimador de máxima verosimilitud, es **multi-modal** y por ende **no-convexo**, es necesario utilizar un algoritmo de **optimización global** para resolverlo.

Posteriormente se extiende dicho resultado, al considerar ceros de **fase no-mínima** para la función de transferencia del ruido $H(q^{-1}, \theta)$ y utilizar el algoritmo de **maximización de la esperanza** (EM) para resolver el problema de estimación.

Mediante ambos enfoques se busca mejorar la precisión del estimador, en términos de varianza del error de estimación, junto con obtener una aproximación razonable de la distribución del ruido $p(w_t)$. Los posibles beneficios del método propuesto se ilustran mediante simulaciones numéricas contrastando los resultados obtenidos con los algoritmos clásicos de identificación [1–3] (como el **método del error de predicción** PEM) y el **método de los momentos** (MoM) [16].

1.2. Estado del arte

Cuando la distribución del ruido w_t en (1.1) es gaussiana, el método del error de predicción (PEM) obtiene el modelo **espectralmente equivalente de fase mínima** (SEMP) (sistema cuyos polos y ceros se encuentran dentro del círculo unitario) del ruido coloreado a la salida del sistema dado por Φ_v [2, 17], con

$$\begin{aligned} y_t &= G(q^{-1}, \theta) u_t + v_t; \\ v_t &= H(q^{-1}, \theta) w_t \\ \Phi_v(z) &= H(z, \theta) H(z^{-1}, \theta) P_w, \end{aligned} \tag{1.3}$$

donde P_w corresponde a la varianza de w_t [18, Capítulo 4]. Esto se debe a que las señales gaussianas quedan completamente determinadas por sus estadísticas de segundo orden [19].

Sin embargo, cuando la distribución del ruido w_t es **no-gaussiana**, es posible identificar transferencias $H(q^{-1}, \theta)$ de fase **no-mínima**, debido a que las estadísticas de alto orden permiten la identificabilidad del sistema $H(q^{-1}, \theta_0)$ (ver [16, 20]).

En [12, 13] se analiza el comportamiento de la **cota inferior de Cramér-Rao** (CRB) de sistemas lineales excitados con ruidos gaussiano y no-gaussiano. En ellos se demuestra que es posible estimar los parámetros del sistema con mayor precisión en el caso no-gaussiano, dado que presenta una CRB inferior que para el caso gaussiano.

Las técnicas presentes en la literatura para estimar sistemas lineales con ruido no-gaussiano explotan dicha característica utilizando el **método de momentos** (MoM) [16, 17, 21–23], el cual requiere un gran número de muestras para producir estimaciones precisas [24], además de restringir la distribución de w_t a cierto conjunto.

Los algoritmos clásicos de identificación, como el **método del error de predicción**, el **método de mínimos cuadrados**, entre otros [1–3], también son frecuentemente utilizados en la práctica, pese a poseer las desventajas señaladas anteriormente.

Problemas que involucran distribuciones no-gaussianas han sido abordados en la literatura, especialmente en el ámbito de **errores en variables** (EIV), el cual trata la problemática de identificar un sistema dinámico cuando todas las variables y señales se encuentran contaminadas con ruido (ver e.g. [25–27]).

El **modelo de mezcla de distribuciones gaussianas** (GMM) ha sido ampliamente utilizado en la literatura para aproximar distribuciones no-gaussianas en diversos ámbitos tales como filtros no-lineales [28, 29], seguimiento [30], comunicaciones [31], estimación bayesiana [32, 33], clasificación [34], reconocimiento de dialectos [35] y estimación de distribuciones [36] por mencionar algunas. En [33] se muestra un enfoque de estimación Bayesiana con métodos hamiltonianos de Monte Carlo (HMC) para sistemas dinámicos con ruido no-gaussiano, no obstante se centra en el caso particular de un modelo auto-regresivo con entrada exógena (ARX) y GMM. Sin embargo, hasta la fecha, no existe una generalización de dicho resultado ni ha sido abordado el problema mediante el método de máxima verosimilitud.

La motivación para considerar sistemas de fase **no-mínima** ha sido previamente considerada en [20, 37–40] entre muchos otros. En [41] se prueba que al obtener muestras de un sistema continua, el modelo de tiempo discreto posee ceros inestables de manera regular y no excepcionalmente.

Si bien la utilización del **error de predicción** (la función de verosimilitud utiliza el error de predicción) en la estimación de los parámetros posee numerosas ventajas [1, 2], también presenta dificultades cuando $H(q^{-1}, \theta)$ posee ceros de fase no-mínima (NMPZ). El predictor requiere de la inversa de H , por ende al poseer NMPZ se vuelve inestable. En [42, 43] se aborda este problema y se presenta una solución mediante filtrado de tiempo reverso.

El algoritmo de **maximización de la esperanza** (EM) [44, 45] es una herramienta para resolver el problema de estimación mediante máxima verosimilitud. Ha sido ampliamente utilizada en identificación de sistemas dinámicos lineales y no-lineales, tanto en el dominio del tiempo (ver e.g. [46–48]) como en el dominio de la frecuencia [49]. La formulación del algoritmo EM con un modelo de mezcla de distribuciones gaussianas GMM ha sido abordado previamente en [36, 50, 51] con resultados satisfactorios.

Entre posibles aplicaciones se encuentran: recuperación de periodicidades ocultas en datos ruidosos [52], estimación de amplitud y fase de sistemas lineales [53, 54], reducción de modelos y determinación de orden de sistemas [16, 55, 56].

Dentro de otras áreas de aplicación destacan: estimación de la fuente en procesa-

miento de datos sísmicos [57,58], construcción de ecualizadores en sistemas de comunicación [59], interferometría de moteado en astronomía [60], procesamiento de imágenes [61], análisis de series de tiempo y predicción de inflación en econometría [62,63], estudios de interacción de oleaje en oceanografía [64], procesamiento de datos astronómicos [65], entre muchas otras.

1.3. Organización de la tesis

A continuación se presenta una breve descripción acerca de los temas tratados en los siguientes capítulos:

- El **Capítulo 2** consiste en una breve introducción a la **identificación de sistemas**. En primer lugar se detallan los tres elementos principales de la identificación de sistemas dados por el conjunto de datos, el conjunto de modelos y un método de identificación. Posteriormente se describen y ejemplifican dos métodos de identificación paramétrica, dados por el método de **máxima verosimilitud** y el **método del error de predicción**, junto a sus propiedades más relevantes. Finalmente se presentan algunas definiciones y propiedades, las cuales serán de esencial ayuda para comprender los análisis realizados en los capítulos posteriores.
- En el **Capítulo 3** se presenta un método presente en la literatura para la estimación de sistemas dinámicos con ruido no-gaussiano, denominado **método de los momentos**. La principal ventaja que posee este enfoque, por sobre el presentado en el **Capítulo 4**, es que considera ceros de **fase no-mínima** en la función de transferencia asociada a las dinámicas del ruido, es decir $H(q^{-1}, \theta)$ (1.1). Inicialmente se presenta una breve introducción y motivación para considerar sistemas de fase no-mínima. Posteriormente se define el problema y los supuestos asociados, para continuar con algunos resultados y definiciones preliminares. Finalmente se detalla el procedimiento del método y algunas consideraciones relevantes.
- El **Capítulo 4** aborda el problema de interés descrito en la **Sección 1.1** restringido a sistemas **fase mínima** para la función de transferencia asociada al ruido, es decir $H(q^{-1}, \theta)$ en (1.1). En primer lugar se analiza el problema de estimación de una constante, cuyas mediciones se encuentran contaminadas con ruido aditivo de distribución **no-gaussiana**, mediante los estimadores de máxima verosimilitud (ML) y mínimos cuadrados (LS). Luego se analiza la **función de verosimilitud** asociada al problema de estimación de una constante en presencia de ruido, cuando la distribución es modelada como una combinación convexa de distribuciones gaussianas (GMM). Posteriormente se desarrolla la expresión que define el estimador de

máxima verosimilitud para sistemas dinámicos, considerando que la distribución del ruido se puede aproximar, o corresponde, a un **modelo de mezcla de distribuciones normales** (GMM). Finalmente se evalúa el desempeño del método propuesto mediante ejemplos numéricos y se analizan los resultados obtenidos.

- El **Capítulo 5** consiste en una extensión del método propuesto en el **Capítulo 4** para la identificación de sistemas lineales, al considerar ceros de **fase no-mínima** para la función de transferencia asociada al ruido $H(q^{-1}, \theta)$ y utilizar el algoritmo de **maximización de la esperanza** (EM) para resolver el problema de estimación mediante máxima verosimilitud. Luego de una breve introducción, se describe el problema y los supuestos asociados. Posteriormente se presenta un desarrollo de una expresión que permite considerar ceros de fase no-mínima en el cálculo del error de predicción, extendiendo así los resultados presentados previamente en los **Capítulos 2 y 4**. Utilizando dicho resultado, se formula el problema de estimación mediante máxima verosimilitud. Después, se desarrolla un algoritmo basado en la maximización de la esperanza para resolver el problema de estimación formulado previamente y se describe el procedimiento utilizado para inicializar el algoritmo. Finalmente se evalúa el desempeño del método propuesto mediante simulaciones numéricas y se analizan los resultados obtenidos. Dichos resultados son contrastados con el método de los momentos MoM (descrito en el **Capítulo 3**) y el método del error de predicción PEM (presentado en la **Sección 2.4**).
- En el **Capítulo 6** se resumen las contribuciones por capítulo. Posteriormente se presentan las conclusiones obtenidas y se proponen posibles trabajos de investigación futuros.
- En el **Apéndice A** se presentan algunas definiciones y resultados de probabilidad y procesos aleatorios utilizados en el desarrollo de este trabajo. Posteriormente se describe el algoritmo de maximización de la esperanza.
- En el **Apéndice B** se introduce al lector en los conceptos y definiciones básicas de **optimización global**. Luego se presentan algunos algoritmos de optimización global heurísticos y uno basado en un patrón de búsqueda (no requiere gradiente). Para cada algoritmo se detalla la estrategia utilizada, el procedimiento a seguir mediante un pseudo-código, algunas consideraciones relevantes y lecturas complementarias. Finalmente se detallan algunas funciones clásicas de prueba, las cuales son utilizadas para evaluar el desempeño de los algoritmos de optimización global. A partir del análisis del desempeño sobre las funciones de prueba, se selecciona un algoritmo, el cual es utilizado para solucionar el problema de optimización asociado a la estimación paramétrica descrita en el **Capítulo 4**.

1.4. Publicaciones realizadas por el autor

Relacionadas a este trabajo

1. Revista científica

- a) R. Orellana, G. Bittner, R. Carvajal and J. C. Agüero, “Maximum Likelihood estimation for non-minimum-phase linear dynamic systems with Gaussian mixture noise distribution,” *Manuscript in preparation to be submitted to Automatica*.

Abstract: In this paper a Maximum Likelihood estimation algorithm for a non-minimum-phase linear dynamic system with Gaussian mixture noise distribution is developed. Based on the Expectation-Maximization algorithm, we propose an identification technique to estimate the system model parameters and the Gaussian mixture parameters. We show that the estimates obtained by using this approach exhibits good accuracy. The benefits of our proposal are illustrated via numerical simulations. ¹

2. Conferencia

- a) G. Bittner, R. Orellana, R. Carvajal and J. C. Agüero, “Maximum Likelihood identification for Linear Dynamic Systems with finite Gaussian mixture noise distribution,” *2019 IEEE CHILEAN Conference on Electrical, Electronics Engineering, Information and Communication Technologies (CHILECON)*, Valparaíso, Chile, 2019, pp. 1-7, DOI: 10.1109/CHILECON47746.2019.8987642.

Abstract: This paper considers the identification of a linear dynamic system driven by a non-Gaussian noise distribution. The noise is approximated by a finite Gaussian mixture, whilst the parameters of the system and the parameters that approximate the noise distribution are simultaneously estimated using the principle of Maximum Likelihood. To this end, a global optimization algorithm is utilized to solve the resulting nonconvex optimization problem. It is shown that our approach improves the accuracy of the estimates, when compared with classic estimation techniques such as the prediction error method (PEM), in terms of covariance of the estimation error, while also obtaining an approximation of the noise distribution. The benefits of the proposed technique are illustrated by numerical simulations ².

¹Presentado parcialmente en el **Capítulo 5**

²Presentado íntegramente en el **Capítulo 4**

Otras

1. Revista científica

- a) Müller N., Kouro S., Zanchetta P., Wheeler P., Bittner G., Girardi F., “Energy Storage Sizing Strategy for Grid-Tied PV Plants under Power Clipping Limitations,” *Energies*, 2019, 12(9):1812, DOI: <https://doi.org/10.3390/en12091812>

Abstract: This paper presents an analyses of an Energy Storage System (ESS) for grid-tied photovoltaic (PV) systems, in order to harness the energy usually lost due to PV array oversizing. A real case of annual PV power generation analysis is presented to illustrate the existing problem and future solutions. Three PV modeling techniques have been applied to estimate non-measured non-harnessed PV power to provide an ESS energy and power sizing strategy. Moreover, a control strategy to store or release power from the DC-link, without modifying the Maximum Power Point Tracking (MPPT) strategy, is presented. The results show an estimation of the annual power loss caused by oversizing the PV array. The ESS sizing strategy gives insight into not only the energy requirements, but also the power requirements of the system. Simulation results show that the proposed ESS control strategy is capable of harnessing the extra power without modifying the existing power converter of the PV plant nor its control strategy.

1.5. Notación

1.5.1. Glosario

$\mathcal{M}$	Conjunto de modelos.
$\mathbb{Z}$	Conjunto de números enteros.
$\mathbb{N}$	Conjunto de números naturales.
$\mathbb{R}$	Conjunto de números reales.
$\mathbb{R}^+$	Conjunto de números reales positivos ($\{x \in \mathbb{R} : x > 0\}$).
Θ	Conjunto de parámetros factibles.
$\mathbb{R}^n$	Conjunto de vectores de dimensión n con elementos reales.
$y_{1:N}$	Conjunto o set de datos con N muestras $y_{1:N} = \{y(t)\}_{t=1}^N$
σ_x	Desviación estándar de la variable aleatoria x ($\text{var}(x) = \sigma_x^2$).
$\mathcal{N}(a, b)$	Distribución normal (gaussiana) con media a y varianza $b \geq 0$.
$p_x \vee p(x)$	Función de densidad de probabilidad (PDF) de la variable aleatoria x .
$P_r\{\cdot\}$	Función de probabilidad.
$\mathcal{L}_N(\beta)$	Función de verosimilitud
$\ell_N(\beta)$	Logaritmo de la función de verosimilitud
$\mathcal{I}_F(\theta)$	Matriz de información de Fisher
μ_x	Media de la variable aleatoria x .
$\text{cov}(x, y)$	Operador covarianza de x e y ($\mathcal{E}\{(x - \mathcal{E}\{x\})(y - \mathcal{E}\{y\})^T\}$). Note que $\text{cov}(x, x) = \text{var}(x)$.
q	Operador desplazamiento en el tiempo ($qu_t = u_{t+1}$).
$\mathcal{E}\{\cdot\}$	Operador esperanza.
$\text{var}(x)$	Operador varianza de x ($\mathcal{E}\{(x - \mathcal{E}\{x\})(x - \mathcal{E}\{x\})^T\}$).
$\hat{\theta}_N$	Parámetro estimado en base al conjunto de datos $\{u(t), y(t)\}_{1:N}$.
θ_0	Parámetro verdadero contenido en Θ (parámetro que describe perfectamente al sistema).
θ	Parámetro(s) utilizado en la estimación.
$\hat{y}_{t t-1}(\theta)$	Predictor a un paso para $y(t)$ dada la colección $\{u(t), y(t)\}_{1:N-1}$.
$x \perp\!\!\!\perp y$	Variables aleatorias x e y independientes.
$x \perp y$	Variables aleatorias x e y no correlacionadas.

1.5.2. Abreviaciones

AIC	Criterio de información de Akaike (Akaike Information Criterion).
a.s.	casi seguro o casi siempre (Almost Surely).
CRB	Cota de Cramér-Rao (Cramér-Rao Bound).
EM	Maximización de la esperanza (Expectation Maximization).
FIM	Matriz de información de Fisher (Fisher Information Matrix).
GA	Algoritmo genético (Genetic Algorithm).
GMM	Modelo de mezcla de distribuciones gaussianas (Gaussian Mixture Model).
i.i.d.	independiente e idénticamente distribuidos (Independent and Identically Distributed).
IIR	Respuesta a impulso infinita (Infinite Impulse Response).
LS	Método de Mínimos Cuadrados (Least Squares).
LTI	Lineal e invariante en el tiempo (Linear Time Invariant).
MAIC	Criterio de información de Akaike modificado (Modified Akaike Information Criterion).
MC	Monte Carlo.
ML	Método de Máxima Verosimilitud (Maximum Likelihood).
MSS	Sentido cuadrático medio (Mean Square Sense).
NMPZ	Ceros de fase no-mínima (Non-Minimum Phase Zeros).
PDF	Función de densidad de probabilidad (Probability Density Function).
PEM	Método del error de predicción (Prediction Error Method).
PS	Optimización mediante búsqueda de patrones (Pattern Search).
PSD	Densidad espectral de potencia (Power Spectral Density).
PSO	Optimización mediante enjambre de partículas (Particle Swarm Optimization).
RS	Muestra aleatoria (Random Sample).
RV	Variable aleatorio (Random Variable).
SA	Recocido simulado (Simulated Annealing).
SE	Espectralmente equivalente (Spectrally Equivalent).
SEMP	Espectralmente equivalente de fase mínima (Spectrally Equivalent Minimum Phase).
SISO	Una entrada y una salida (Single Input Single Output).
w.p.1.	con probabilidad uno (With Probability 1).

1.5.3. Convenciones de Notación

$H^{-1}(q^{-1})$	$[H(q^{-1})]^{-1}$
x_t^T	$[x_t]^T$
A^{-T}	$[A^{-1}]^T$
$\triangleq$	Definido por/como.
$\sim$	Distribuido según.

1.5.4. Convenciones de Polinomios

$A(q)$	$q^n + a_1 q^{n-1} + \dots + a_n$
$A(q^{-1})$	$q^{-n} + a_1 q^{-(n-1)} + \dots + a_n$

IDENTIFICACIÓN DE SISTEMAS

La obtención de **modelos matemáticos**, o modelado, representa una fase crucial en diversas áreas y aplicaciones. Por modelado se entiende al proceso de obtener una descripción matemática de un fenómeno, al cual llamaremos **sistema**. Ejemplos incluyen pronósticos de la demanda de electricidad [66], pronósticos bursátiles [67], estimación del canal de comunicaciones [68], predicciones de migración aviar [69], modelos de propagación de enfermedades [70, 71], diseño de controladores en procesos industriales [1–3], entre muchos otros.

Existen dos enfoques principales al modelado. El primer enfoque considera el modelado del sistema o fenómeno a partir de leyes físicas, por ejemplo el uso de las leyes de Kirchhoff de tensión y corriente para modelar un circuito eléctrico [72].

El segundo enfoque considera el modelado del sistema como un modelo de caja negra (black box), o caja gris (gray box) en caso de contar con información acerca de las leyes físicas del sistema, e identificar el modelo en base a datos de entrada y salida obtenidos a partir de un experimento realizado sobre el sistema (e.g. aplicar una tensión sobre un circuito eléctrico y medir la corriente para obtener un modelo para la impedancia equivalente). Este trabajo se centra en el segundo enfoque, el cual se conoce comúnmente como **identificación de sistemas**.

A continuación se presentan algunos antecedentes relevantes respecto de la identificación de sistemas, lo cual será de suma utilidad en los capítulos posteriores.

Observación 2.1. *En caso de encontrarse familiarizado con identificación de sistemas, se sugiere continuar en el **Capítulo 3**, donde se presenta el método de los momentos MoM.*

2.1. Sistema

La identificación de sistemas concierne el modelado de un sistema en base a datos de entrada y salida obtenidos de él. Algunos autores [1, 2] han definido **sistema** como *un*

objeto en el cual variables de diferentes tipos interactúan para producir señales observables. En este contexto, se denominan **entradas** a los estímulos externos que afectan el sistema, los cuales pueden ser manipulados por el usuario, y **salidas** a las señales observables de interés. Adicionalmente, el sistema puede estar sujeto a **perturbaciones**, las cuales corresponden a estímulos que no pueden ser controlados (no obstante algunas veces pueden ser medibles). En la Fig. 2.1 se puede apreciar un diagrama de bloques que representa el concepto de sistema. En dicha figura u_t corresponde a la entrada, y_t a la salida y v_t a las perturbaciones.

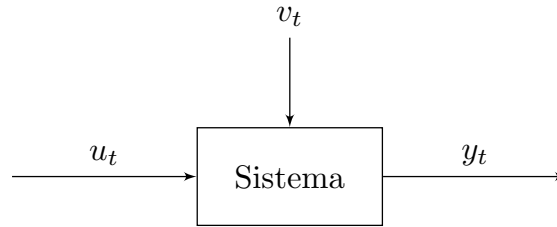


Fig. 2.1: Diagrama de bloques de un sistema genérico.

A partir de la definición de sistema se puede inferir que la teoría de identificación de sistemas puede ser aplicada en numerosas áreas y problemas. A continuación se describen los elementos principales de la teoría de identificación de sistemas.

2.2. Elementos principales de identificación de sistemas

Para construir un modelo a partir de datos se requieren 3 entidades básicas. En primer lugar es necesario un arreglo o **conjunto de datos** de entrada-salida, un **conjunto de modelos** y un **método de identificación** que permita encontrar un modelo (adecuado en cierto sentido) a partir del conjunto de candidatos. A continuación revisaremos brevemente dichos elementos

2.2.1. Conjunto de Datos

Para poder estimar un modelo es necesario contar con un arreglo de datos de entrada-salida del sistema, el cual denominaremos **conjunto de datos**. De esta manera y siguiendo la notación utilizada en la Fig. 2.1, se define el conjunto de datos como

$$\{y_t, u_t\}_{t=1}^N, \quad (2.1)$$

donde N representa el número de datos.

2.2.2. Conjunto de Modelos

A partir del conjunto de datos $\{y_t, u_t\}_{t=1}^N$ se desean estimar los parámetros de un modelo para el sistema. Dicho modelo se encuentra usualmente restringido a un conjunto de descripciones matemáticas posibles denominado **conjunto de modelos** $\mathcal{M}$.

Cabe destacar que los modelos contenidos en $\mathcal{M}$ se encuentran parametrizados por $\theta \in \Theta$, donde θ y Θ corresponden a los parámetros utilizados en la estimación y al conjunto de parámetros factibles respectivamente.

La elección de dicho conjunto se realiza en base a la información previa disponible del sistema, utilizando técnicas no-paramétricas o una combinación de ambas [1–3].

Observación 2.2 (Submodelado).

*Un supuesto ampliamente utilizado sobre el conjunto de modelos $\mathcal{M}(\theta)$ es que existe al menos un $\theta_0 \in \Theta$ tal que al evaluar el modelo en θ_0 describe perfectamente al sistema. Si dicha condición no se satisface, entonces es posible afirmar que el sistema se encuentra **sub-modelado** por $\mathcal{M}(\theta)$.* ∇

Existen diversos conjuntos de modelos descritos en la literatura [1–3, 27, 73–75], no obstante en este trabajo nos centraremos en una clase específica de modelos, denominada **sistemas dinámicos lineales e invariantes en el tiempo**, la cual es presentada a continuación.

Sistemas lineales invariantes en el tiempo

Si el sistema opera linealmente en torno a un punto de operación [72, 76] resulta razonable considerar que el sistema puede encontrarse representado por un modelo dentro del conjunto definido por

$$\mathcal{M}(\theta) \triangleq \{y_t = G(q; \theta) u_t + H(q; \theta) \omega_t \mid \theta \in \Theta\}, \quad (2.2)$$

donde $\Theta \in \mathbb{R}^n$ (vector de dimensión n con coeficientes reales) corresponde al conjunto de parámetros factibles, $G(q; \theta)$ y $H(q; \theta)$ corresponden a funciones racionales del operador desplazamiento en el tiempo q ($qu_t = u_{t+1}$) parametrizadas por $\theta \in \Theta$ y ω_t representa una secuencia de ruido blanco de media cero ($\mu(\omega_t) = 0$) y varianza finita ($P(\omega_t) < \infty$).

Dependiendo de los supuestos realizados sobre $G(q; \theta)$ y $H(q; \theta)$ es posible describir el sistema mediante diversas estructuras. Dichas estructuras se pueden clasificar (ver Tabla 2.1) dependiendo de los polinomios utilizados en las definiciones de $G(q; \theta)$ y $H(q; \theta)$ [1, 2].

Acrónimo	Nombre	Estructura
AR	Autoregresivo	$y_t = \frac{1}{A(q; \theta)} \omega_t$
MA	Media Móvil	$y_t = C(q; \theta) \omega_t$
ARMA	Autoregresivo de Media Móvil	$y_t = \frac{C(q; \theta)}{A(q; \theta)} \omega_t$
FIR	Respuesta Finita al Impulso	$y_t = B(q; \theta) u_t + \omega_t$
ARX	Autoregresivo con entrada Exógena	$y_t = \frac{B(q; \theta)}{A(q; \theta)} u_t + \frac{1}{A(q; \theta)} \omega_t$
OE	Error de Salida	$y_t = \frac{B(q; \theta)}{A(q; \theta)} u_t + \omega_t$
B-J	Box-Jenkins	$y_t = \frac{B(q; \theta)}{A(q; \theta)} u_t + \frac{C(q; \theta)}{D(q; \theta)} \omega_t$

Tabla 2.1: Algunas clasificaciones de los sistemas LTI de acuerdo a su estructura.

Observación 2.3. *Algunos autores incluyen un polinomio adicional al lado izquierdo de algunas estructuras, por ejemplo B-J, no obstante dicho polinomio puede ser omitido sin pérdida de generalidad.* ∇

Observación 2.4. *La perturbación v_t presente en la Fig. 2.1 puede ser descrita de la siguiente manera*

$$v_t = H(q; \theta_0) \omega_t \quad (2.3)$$

Para algún $\theta_0 \in \Theta$. Para mayor información se sugiere revisar **Factorización Espectral** [18, Capítulo 4]. ∇

2.2.3. Método de Identificación

El objetivo es encontrar un modelo contenido en $\mathcal{M}(\theta)$ que explique de mejor manera los datos $\{u_t, y_t\}_{t=1:N}$ en algún sentido (dado por un funcional de costo). De manera más formal, se debe determinar un mapeo desde los datos contenidos en $\{u_t, y_t\}_{t=1:N}$ hacia el conjunto de modelos $\mathcal{M}$ parametrizados por θ , es decir

$$\{u_t, y_t\}_{t=1:N} \rightarrow \mathcal{M}(\theta) \in \mathcal{M}(\Theta) \quad (2.4)$$

Dicho mapeo se realiza mediante un algoritmo o método, el cual se denomina **método de identificación** (particularmente, **método de estimación paramétrica** [1–3]).

Si bien existen numerosos métodos de identificación en la literatura, por ejemplo **mínimos cuadrados** (LS) [2, Capítulo 4], **variables instrumentales** (IV) [77], nos centraremos en los siguientes dos métodos de identificación

1. **Método de máxima verosimilitud** (ML)
2. **Método de error de predicción** (PEM)

Dichos métodos se detallan en las siguientes secciones.

2.3. Método de máxima verosimilitud (ML)

El **método de máxima verosimilitud** (ML) es uno de los métodos de identificación de sistemas más atractivos debido a sus propiedades estadísticas. El método de máxima verosimilitud se basa en la *distribución del conjunto de datos* o muestra aleatoria $y_{1:N}$ (las observaciones son modeladas como realizaciones de variables aleatorias, ver **Definición 2.1**), la cual se encuentra parametrizada por $\theta \in \Theta$. El objetivo es encontrar el parámetro estimado $\hat{\theta}_{ML}$ que explique de mejor manera las muestras $y_{1:N}$. Al definir $p(y_{1:N}|\theta)$ como la función de densidad de probabilidad (PDF) conjunta asociada a los datos $y_{1:N}$, es posible estimar $\hat{\theta}_N$ mediante el método de máxima verosimilitud como

$$\hat{\theta}_{ML} \triangleq \arg \max_{\theta \in \Theta} p(y_{1:N}|\theta) \quad (2.5)$$

La expresión dada por (2.5) se puede interpretar intuitivamente como: $\hat{\theta}_N \in \Theta$ corresponde al parámetro que hace a los datos observados $y_{1:N}$ más probables [1, 2].

La función $\mathcal{L}_N(\theta) = p(y_{1:N}, \theta)$ se denomina generalmente **verosimilitud de los datos** o **función de verosimilitud**.

Observación 2.5. *Hallar $\hat{\theta}_{ML}$ es, en general, muy difícil (salvo para el caso lineal y gaussiano) dado que la distribución puede ser muy compleja o difícil de parametrizar. En general se utilizan algoritmos numéricos para aproximar $\hat{\theta}_{ML}$. Ver **algoritmo de maximización de la esperanza** (EM) en la **Sección A.2**. [44, 78].* ∇

Usualmente es preferible maximizar el logaritmo de la función de densidad de probabilidad, es decir $\ell_N(\theta) \triangleq \ln p(y_{1:N}|\theta)$, en vez de $p(y_{1:N}|\theta)$. Dicha expresión ($\ell_N(\theta)$) se conoce como **logaritmo de la función de verosimilitud** (o *log-likelihood function*). Dado que la función logaritmo es estrictamente creciente, entonces la solución de

$$\hat{\theta}_{ML} \triangleq \arg \max_{\theta \in \Theta} \ln p(y_{1:N}|\theta) \quad (2.6)$$

es equivalente solución (2.5). En general se utiliza (2.6) por sobre (2.5) puesto que los productos contenidos en $p(y_{1:N}|\theta)$ se transforman en sumas y debido a que la función logaritmo remueve las exponenciales (lo cual resulta favorable cuando la función de distribución de probabilidad (PDF) es del tipo exponencial ya que $\ln e^a = a$). Otra razón por la cual se prefiere utilizar el logaritmo de la función de verosimilitud es debido a que dan lugar a algoritmos mejor condicionados numéricamente [79, 80].

Ejemplo 2.1. Considere una muestra aleatoria $x_{1:N}$ de la variable aleatoria escalar $X \sim \mathcal{N}(\mu_x, \sigma_x^2)$. Encuentre estimadores de máxima verosimilitud para μ_x y σ_x^2 .

Dado que se trata de una muestra aleatoria, es posible caracterizar la función de probabilidad conjunta como

$$p(x_{1:N}|\theta) = \prod_{i=1}^N \frac{1}{\sqrt{2\pi\sigma_x^2}} \exp\left\{-\frac{(x_i - \mu_x)^2}{2\sigma_x^2}\right\} \quad (2.7)$$

con $\theta \triangleq [\mu_x \quad \sigma_x^2]^T$. Reemplazando (2.7) en (2.6) es posible obtener

$$\begin{aligned} \hat{\theta}_{ML} = \hat{\theta}_N &= \arg \max_{\theta \in \mathbb{R}^2} \ln p(x_{1:N}|\theta) \\ &= \arg \max_{\theta \in \mathbb{R}^2} \sum_{i=1}^N \left(-\frac{1}{2} \ln(2\pi\sigma_x^2) - \frac{1}{2} \frac{(x_i - \mu_x)^2}{\sigma_x^2} \right) \\ &= \arg \max_{\theta \in \mathbb{R}^2} \ell_N(\theta) \quad ; \ell_N(\theta) = \left(-\frac{N}{2} \ln(2\pi\sigma_x^2) - \sum_{i=1}^N \frac{1}{2\sigma_x^2} (x_i - \mu_x)^2 \right) \end{aligned} \quad (2.8)$$

A partir de $\ell_N(\theta)$ es posible obtener el estimador para la media de x como

$$\begin{aligned} \frac{\partial \ell_N(\theta)}{\partial \theta_1} &= \frac{\partial \ell_N(\theta)}{\partial \mu_x} = -\frac{1}{2\sigma_x^2} \sum_{i=1}^N 2(x_i - \mu_x)(-1) \\ &= \frac{1}{\sigma_x^2} \sum_{i=1}^N (x_i - \mu_x) = 0 \end{aligned} \quad (2.9)$$

Despejando se tiene

$$\hat{\mu}_{x_{ML}} = \frac{1}{N} \sum_{i=1}^N x_i \quad (2.10)$$

De manera análoga es posible obtener el estimador para la varianza de x

$$\frac{\partial \ell_N(\theta)}{\partial \theta_2} = \frac{\partial \ell_N(\theta)}{\partial \sigma_x^2} = -\frac{N}{2} \frac{1}{2\pi\sigma_x^2} 2\pi + \frac{1}{2\sigma_x^4} \sum_{i=1}^N (x_i - \mu_x)^2 = 0 \quad (2.11)$$

Despejando se obtiene

$$\hat{\sigma}_{x_{ML}}^2 = \frac{1}{N} \sum_{i=1}^N (x_i - \hat{\mu}_{x_{ML}})^2 \quad (2.12)$$

Los estimadores obtenidos son ampliamente conocidos y es posible demostrar que $\hat{\mu}_{x_{ML}}$ es *insesgado, eficiente y consistente*. Del mismo modo es posible demostrar que $\hat{\sigma}_{x_{ML}}^2$ es *asintóticamente insesgado, asintóticamente eficiente y consistente* (ver e.g. [81]). Estas propiedades estadísticas son típicas de los estimadores de máxima verosimilitud, por lo cual resultan particularmente atractivos en identificación.

2.3.1. Análisis asintótico

Los estimadores obtenidos mediante el **método de máxima verosimilitud** ML y el **método del error de predicción** PEM (para un caso particular de $l(\cdot)$ e innovaciones Gaussianas, el cual se analiza posteriormente en el **Ejemplo 2.2**) son asintóticamente eficientes. Entonces, resulta interesante analizar como el estimador $\hat{\theta}_N$ se comporta cuando $N \rightarrow \infty$. El resultado clave, introducido en [82, 83], acerca de la distribución asintótica de los estimadores de ML al modelar las observaciones como realizaciones de variables aleatorias i.i.d. se presenta a continuación

Lema 2.1 (Distribución asintótica del estimador de máxima verosimilitud).

Suponga que las variables aleatorias (RV) $x_{1:N}$ son independientes e idénticamente distribuidas (i.i.d.). Adicionalmente asuma que la distribución de $x_{1:N}$ esta dada por $p(x_{1:N}|\theta_0)$ para algún $\theta_0 \in \Theta$. Entonces cuando $N \rightarrow \infty$, la variable aleatoria $\hat{\theta}_N$ tiende a θ_0 con probabilidad uno (w.p.1.), y la variable aleatoria $\sqrt{N}(\hat{\theta}_N - \theta_0)$ converge en distribución hacia

$$\sqrt{N}(\hat{\theta}_N - \theta_0) \sim As \mathcal{N}(0, \mathcal{I}_F^{-1}(\theta)), \quad (2.13)$$

donde $\mathcal{I}_F(\theta)$ se encuentra dado por (2.36) (ver **Teorema 2.2**).

Demostración 2.1. Ver [1, Capítulo 9 y Apéndice 9]. □

El **Lema 2.1** indica que, cuando $N \rightarrow \infty$, la distribución de $\sqrt{N}(\hat{\theta}_N - \theta_0)$ tiende a una normal con media cero y matriz de covarianza dada por la cota inferior de Cramér-Rao. Este resultado también es valido bajo ciertas condiciones leves [1] cuando se aplica el método de máxima verosimilitud en sistemas dinámicos. El **Lema 2.1** se aplica también al estimador obtenido mediante el método del error de predicción PEM cuando la secuencia de ruido blanco dada por ω_t es Gaussiana y la función $l(\cdot)$ es una función cuadrática [1].

Corolario 2.1.

A partir de lo anteriormente expuesto se tiene que (ver [1, 83]):

- $\hat{\theta}_{ML}$ es asintóticamente insesgado.
- $\hat{\theta}_{ML}$ es asintóticamente eficiente.
- $\hat{\theta}_{ML}$ es consistente.

2.4. Método del error de predicción (PEM)

El **método del error de predicción** (PEM) es otra manera de abordar el problema de encontrar el mejor modelo a partir del conjunto de modelos $\mathcal{M}(\theta)$ que explique los datos contenidos en $\{u_t, y_t\}_{t=1:N}$ (ver e.g. [1, 84, 85]). La idea básica del método del error de predicción es bastante simple y consta de los siguientes pasos:

1. Describa el modelo como un **predictor** de la siguiente salida:

$$\hat{y}_{t|t-1} = f(\{u_i, y_i\}_{i=1:t-1}) \quad (2.14)$$

Donde $\hat{y}_{t|t-1}$ representa la predicción a un paso de la salida y_t , y $f(\cdot)$ corresponde a una función arbitraria de los datos pasados (hasta el instante $t-1$).

2. Parametrice el predictor en términos de un vector de parámetros de dimensión finita denotado por $\theta \in \Theta$:

$$\hat{y}_{t|t-1}(\theta) = f(\{u_i, y_i\}_{i=1:t-1}, \theta) \quad (2.15)$$

Algunas condiciones de regularidad son necesarias para realizar dicha parametrización. Para mayor información se sugiere revisar [1, Capítulo 4].

3. Defina el **error de predicción** como la distancia entre la predicción $\hat{y}_{t|t-1}(\theta)$ y la salida $y(t)$:

$$\varepsilon_t(\theta) = y_t - \hat{y}_{t|t-1}(\theta) \quad (2.16)$$

Opcionalmente, es posible filtrar la secuencia de error de predicción mediante un filtro LTI estable arbitrario $L(q)$ antes de la estimación [1, Capítulo 7]:

$$\varepsilon_t^F(\theta) = L(q)\varepsilon_t(\theta) \quad (2.17)$$

4. Estime θ (denotada por $\hat{\theta}$) a partir de la parametrización y del conjunto de datos $\{u_i, y_i\}_{i=1:N}$ minimizando la distancia entre $\hat{y}_1(\theta), \dots, \hat{y}_N(\theta)$ y $y_1, \dots, y_N$ en alguna norma, es decir

$$\hat{\theta}_{PEM} = \hat{\theta}_N = \arg \min_{\theta \in \Theta} V_N(\theta) \quad (2.18)$$

Con

$$V_N(\theta) = \sum_{t=1}^N l(\varepsilon_t^F(\theta)) \quad (2.19)$$

Donde $l(\cdot)$ representa una función positiva arbitraria definida por el usuario (generalmente se utiliza una norma apropiada).

Cabe destacar que existen varios métodos de identificación contenidos dentro de PEM dependiendo de las elecciones de $l(\cdot)$, $L(q)$, estructura de los modelos $\mathcal{M}(\theta)$ y métodos de minimización.

La minimización del error de predicción $\varepsilon_t(\theta)$ tiene sentido, puesto que los modelos obtenidos son utilizados tanto en predicción como en síntesis de controladores.

Note que la función $V_N(\theta)$ corresponde a una función escalar bien definida del parámetro θ . Dicha función puede ser considerada como una medida natural de la validez del modelo $\mathcal{M}(\theta)$.

Observación 2.6. Si la función (o norma) previamente mencionada $l(\cdot)$ es elegida adecuadamente, la estimación $\hat{\theta}_N$ coincide con la estimación mediante el Método de Máxima Verosimilitud. Ver **Ejemplo 2.2** (para mayor información se sugiere revisar [1, Capítulo 7] y [86]). ∇

El método de error de predicción (PEM) posee algunas ventajas [1, 84, 85], entre las cuales destacan:

- Puede ser aplicado a un amplio espectro de parametrizaciones del modelo puesto que solo requiere una expresión para (2.15).
- Los modelos obtenidos poseen excelentes propiedades asintóticas debido a la cercanía entre PEM y ML (Ver **Ejemplo 2.2**).
- Soporta sistemas operando en lazo cerrado (es decir, la entrada del sistema se encuentra parcialmente determinada por realimentación de la salida al momento de recolectar los datos) sin la necesidad de recurrir a técnicas adicionales. Esta característica también es válida para el método de máxima verosimilitud.

No obstante, también posee algunos inconvenientes entre los cuales destacan [1, Capítulo 7]:

- Requiere una parametrización explícita del modelo (e.g. un modelo ARX de quinto orden (ver Tabla 2.1)).
- La búsqueda de los parámetros que maximizan la predicción de la salida puede ser tediosa e involucrar una búsqueda sobre superficies con múltiples mínimos locales.
- El predictor a un paso de la salida $\hat{y}_{t|t-1}(\theta)$ debe ser estable.

La conexión entre PEM y ML (ver **Observación 2.6**) se puede apreciar a continuación:

Ejemplo 2.2 (Relación entre PEM y ML).

Considere un sistema SISO que puede ser expresado mediante

$$y_t = \hat{y}_{t|t-1}(\theta_0) + \omega_t, \quad (2.20)$$

Para algún θ_0 . Con ω_t ruido blanco Gaussiano de media cero $\mu\omega_t = 0$ y varianza $\sigma^2(\omega_t)(\theta_0)$ desconocida. Si el predictor $\hat{y}_{t|t-1}(\theta_0)$ se encuentra dado por

$$\hat{y}_{t|t-1}(\theta_0) = \mathcal{E}\{y_t | y_{1:t-1}\} \quad (2.21)$$

Entonces, bajo las condiciones anteriores, es posible obtener el logaritmo de la función de verosimilitud como

$$\ell_N(\theta) = \ln p(y_{1:N} | \theta) = -\frac{1}{2} \sum_{t=1}^N \varepsilon_t^2(\theta) - \frac{N}{2} \ln \sigma^2(\theta) + c \quad (2.22)$$

Con ε_t dado por (2.16) y c una constante independiente de θ . Al asumir que $\varepsilon_t(\theta)$ y $\sigma^2(\theta)$ se encuentran independientemente parametrizados por θ , entonces es posible maximizar dicha expresión sobre $\sigma^2(\theta)$ para obtener una expresión en función de $\varepsilon_t(\theta)$. Esto permite reescribir (2.22) como

$$\ln p(y_{1:N} | \theta) = -\frac{1}{N} \sum_{t=1}^N \varepsilon_t^2(\theta) + c \quad (2.23)$$

Al comparar el funcional de PEM dado por (2.19) con (2.23) es posible apreciar que el **método del error de predicción** deriva del **método de máxima verosimilitud** cuando la secuencia $\{\omega_t\}$ es ruido blanco gaussiano con media cero y se elige $l(\varepsilon_t(\theta)) = N^{-1} \varepsilon_t^2(\theta)$.

No obstante el predictor definido por (2.21) (el cual corresponde al estimador óptimo en sentido cuadrático medio a un paso de y_t dada la colección de datos $y_{1:t-1}$ [18, Capítulo 5]) no posee una expresión cerrada en general.

Exceptuando el caso lineal (ver sistemas lineales e invariantes en el tiempo en la **Subsección 2.2.2**) y casos particulares de sistemas no-lineales, la expresión (2.21) solo puede ser obtenida de manera numérica, utilizando, por ejemplo, el método de **filtro de partículas** [87].

A continuación es posible apreciar la expresión cerrada para (2.21) utilizada para sistemas lineales.

Ejemplo 2.3. Considere el conjunto de modelos $\mathcal{M}(\theta)$ definido por (2.2). A continuación se obtiene el predictor a un paso $\hat{y}_{t|t-1}(\theta)$ con $\theta \in \Theta$. Suponga que $\lim_{q \rightarrow \infty} H(q; \theta) = I$. Bajo este supuesto es posible reescribir cualquier modelo a partir de (2.2) (ver Tabla 2.1) como

$$y_t = G(q; \theta) u_t + (H(q; \theta) - I) \omega_t + \omega_t \quad (2.24)$$

Note que el término $(H(q; \theta) - I) \omega_t$ en (2.24) contiene solo términos hasta $t-1$. A partir de (2.2) es posible despejar ω_t en función de y_t y u_t e insertar dicho resultado en (2.24) obteniendo de esta manera

$$y_t = H^{-1}(q; \theta) G(q; \theta) u_t + (I - H^{-1}(q; \theta)) y_t + \omega_t \quad (2.25)$$

Note que los términos $H^{-1}(q; \theta) G(q; \theta) u_t$ y $(I - H^{-1}(q; \theta)) y_t$ solo dependen de $\{y_i, u_i\}_{i=1}^{t-1}$ si $G(q; \theta)$ es estrictamente propia (si u_t es determinística no es necesaria dicha condición). Adicionalmente dado que $\{\omega_t\}$ corresponde a una secuencia de ruido blanco, se tiene que la mejor estimación de ω_t dado $\{y_i, u_i\}_{i=1}^{t-1}$ es $\mathcal{E}\{\omega_t\} = 0$. De esta manera el predictor a un paso asociado al conjunto de modelos (2.2) se encuentra dado por

$$\hat{y}_{t|t-1}(\theta) = H^{-1}(q; \theta) G(q; \theta) u_t + (I - H^{-1}(q; \theta)) y_t \quad (2.26)$$

La expresión (2.26) corresponde a un predictor válido para y_t si $H^{-1}(q; \theta) G(q; \theta)$ y $H^{-1}(q; \theta)$ son estables [1, Capítulo 5].

Observación 2.7. En el **Ejemplo 2.3**, ω_t no puede ser estimado a partir de $\{y_i, u_i\}_{i=1}^{t-1}$ puesto que ω_t es ruido blanco. Dada esta propiedad, se denomina ω_t como un proceso de innovación [18, Capítulo 6]. ∇

2.5. Algunas definiciones

A continuación se presentan algunas propiedades y definiciones relativas a los estimadores paramétricos

Definición 2.1 (Muestra aleatoria).

Dada una variable aleatoria (RV) X , se define una **muestra aleatoria** (RS) de X como una secuencia finita i.i.d. (independiente e idénticamente distribuidas) de variables que distribuyen como X . Para ello utilizaremos la siguiente notación

$$\{x_i : i = 1, \dots, N\} \quad o \quad \{x_{1:N}\} \quad (2.27)$$

Definición 2.2 (Sentido cuadrático medio).

Considere un estimador de y dado por $\hat{y}(\theta)$. Dicho estimador es óptimo en **sentido cuadrático medio** (MSS) ssi

$$\hat{y}(\theta) = \arg \min_{\theta \in \Theta} J(\theta) \quad (2.28)$$

Donde $J(\theta)$ corresponde al funcional definido por

$$J(\theta) \triangleq \mathcal{E} \left\{ (y - \hat{y}(\theta)) (y - \hat{y}(\theta))^T \right\} \quad (2.29)$$

Note que es posible definir el **error de estimación** como $\varepsilon = y - \hat{y}(\theta)$, por ende el estimador óptimo en sentido cuadrático medio reduce la esperanza del error de estimación al cuadrado [18].

Definición 2.3 (Sesgo).

Sea $\hat{\theta}$ un estimador del parámetro θ basado en una muestra aleatoria (RS) de dimensión N . Entonces

- El estimador $\hat{\theta}$ es **insesgado** ssi

$$\mathcal{E} \left\{ \hat{\theta} \right\} = \theta \quad ; \forall \theta \quad (2.30)$$

- $\hat{\theta}$ es **asintóticamente insesgado** ssi

$$\lim_{N \rightarrow \infty} \mathcal{E} \left\{ \hat{\theta} \right\} = \theta \quad (2.31)$$

Dichas definiciones tiene sentido puesto que $\hat{\theta}$ es una variable aleatoria (es una estadística de una muestra aleatoria).

Que un estimador sea insesgado no es particularmente útil en la práctica, puesto que para poder recuperar θ en base a un estimador insesgado es necesario repetir el experimento que genera la muestra aleatoria (si el número de experimentos es suficientemente grande es posible utilizar la ley de los grandes números para aproximar la esperanza).

Definición 2.4 (Consistencia).

Sea $\hat{\theta}$ un estimador del parámetro θ basado en una muestra aleatoria (RS) de dimensión N .

- $\hat{\theta}$ es **consistente** ssi

$$\lim_{N \rightarrow \infty} \Pr \left\{ \left| \hat{\theta} - \theta \right| > \epsilon \right\} = 0 \quad ; \forall \epsilon > 0 ; \forall \theta \quad (2.32)$$

Un estimador es consistente si a medida que el tamaño de la muestra aleatoria N crece, la estimación θ tiende al parámetro θ_0 . Note que para explotar esta propiedad es necesario extender el largo del experimento y no repetirlo, lo cual resulta particularmente útil en la práctica.

Para mayor información y detalle acerca de las propiedades de sesgo y consistencia de los estimadores paramétricos se sugiere revisar [81, Capítulos 7 y 8].

Teorema 2.1. Sea $\hat{\theta}$ un estimador para el parámetro θ basado en una muestra aleatoria de dimensión N . Entonces, si

$$\lim_{N \rightarrow \infty} \mathcal{E} \left\{ \hat{\theta} \right\} = \theta \quad (2.33)$$

y

$$\lim_{N \rightarrow \infty} \text{var}(\hat{\theta}) = \lim_{N \rightarrow \infty} P_{\hat{\theta}} = \lim_{N \rightarrow \infty} \mathcal{E} \left\{ \left(\hat{\theta} - \mathcal{E} \left\{ \hat{\theta} \right\} \right) \left(\hat{\theta} - \mathcal{E} \left\{ \hat{\theta} \right\} \right)^T \right\} = 0 \quad (2.34)$$

entonces el estimador $\hat{\theta}$ es consistente. En (2.34) $P_{\hat{\theta}}$ corresponde a la varianza del error de estimación.

Note que un estimador puede tener sesgo y aún así ser consistente (este resultado exige que $\hat{\theta}$ sea asintóticamente insesgado).

Demostración 2.2. Ver [81, Capítulos 7 y 8] y [18]. □

Teorema 2.2 (Cota de Cramér-Rao).

Suponga que la función de densidad de probabilidad (PDF) de una variable aleatoria Y se

encuentra parametrizada por θ , es decir $p(y|\theta)$. Considere ahora una muestra aleatoria de tamaño N de la variable aleatoria $(y_{1:N})$. Todo estimador sin sesgo para θ_0 satisface

$$\mathcal{E} \left\{ \left(\hat{\theta}_N - \theta_0 \right) \left(\hat{\theta}_N - \theta_0 \right)^T \right\} \geq \mathcal{I}_F^{-1}(\theta) \quad (2.35)$$

Donde

$$\mathcal{I}_F(\theta) \triangleq \mathcal{E} \left\{ \left[\frac{\partial L(y_{1:N}|\theta)}{\partial \theta} \Big|_{\theta=\theta_0} \right] \left[\frac{\partial L(y_{1:N}|\theta)}{\partial \theta} \Big|_{\theta=\theta_0} \right]^T \right\} \quad (2.36)$$

con

$$\begin{aligned} \frac{\partial(\cdot)}{\partial \theta} &\triangleq \begin{bmatrix} \frac{\partial(\cdot)}{\partial \theta_1} & \dots & \frac{\partial(\cdot)}{\partial \theta_n} \end{bmatrix} \\ L(y_{1:N}|\theta) &\triangleq \ln p(y_{1:N}|\theta) \end{aligned} \quad (2.37)$$

$\mathcal{I}_F(\theta)$ se denomina **matriz de información de Fisher (FIM)**, mientras que $\mathcal{I}_F^{-1}(\theta)$ se denomina **cota inferior de Cramér-Rao (CRB)** [83] [1, Capítulo 7].

Observación 2.8. La matriz de información de Fisher $\mathcal{I}_F(\theta)$ depende únicamente de la muestra aleatoria (y de su distribución), por ende $\mathcal{I}_F(\theta)$ cuantifica la información de θ_0 contenida en $y_{1:N}$.

El cálculo de $\mathcal{I}_F(\theta)$ requiere conocer el valor de θ_0 , por ende generalmente no se dispone del valor exacto de la matriz de información de Fisher.

Dado que este resultado (2.35) y (2.36) es válido para todo estimador, la cota de Cramér-Rao (CRB) establece un límite fundamental (tanto para estimadores lineales como no-lineales). ∇

Demostración 2.3 (Cota de Cramér-Rao).

Considere el caso en el cual $\theta \in \Theta = \mathbb{R}$ (es decir hay que estimar un solo parámetro).

Sea $\hat{\theta}_N$ un estimador insesgado de θ basado en la muestra aleatoria $x_{1:N}$ dado por

$$\hat{\theta}_N = g(x_{1:N}) \quad (2.38)$$

Puesto que $\hat{\theta}_N$ es insesgado, se tiene que $\mathcal{E} \left\{ \hat{\theta}_N \right\} = \theta$, lo cual puede reescribirse como

$$\begin{aligned} \theta &= \int_{-\infty}^{\infty} g(x_{1:N}) p(x_{1:N}|\theta) dx_1^N \\ 1 &= \frac{\partial}{\partial \theta} \int_{-\infty}^{\infty} g(x_{1:N}) p(x_{1:N}|\theta) dx_1^N \\ 1 &= \int_{-\infty}^{\infty} g(x_{1:N}) \frac{\partial p(x_{1:N}|\theta)}{\partial \theta} dx_1^N \end{aligned}$$

$$\begin{aligned}
1 &= \int_{-\infty}^{\infty} g(x_{1:N}) \frac{\partial L(x_{1:N}|\theta)}{\partial \theta} p(x_{1:N}|\theta) dx_1^N \\
1 &= \mathcal{E} \left\{ g(x_{1:N}) \frac{\partial L(x_{1:N}|\theta)}{\partial \theta} \right\}
\end{aligned} \tag{2.39}$$

Por otra parte se tiene que

$$\begin{aligned}
1 &= \int_{-\infty}^{\infty} p(x_{1:N}|\theta) dx_1^N \\
0 &= \int_{-\infty}^{\infty} \frac{\partial L(x_{1:N}|\theta)}{\partial \theta} p(x_{1:N}|\theta) dx_1^N \\
0 &\stackrel{(a)}{=} \mathcal{E} \left\{ \theta \frac{\partial L(x_{1:N}|\theta)}{\partial \theta} \right\}
\end{aligned} \tag{2.40}$$

En (a) es posible agregar θ dentro del operador esperanza, dado que $\theta \in \Theta = \mathbb{R}$. Utilizando (2.39) y (2.40) es posible obtener

$$\begin{aligned}
1 &= \mathcal{E} \left\{ (g(x_{1:N}) - \theta) \frac{\partial L(x_{1:N}|\theta)}{\partial \theta} \right\} \\
1 &= \mathcal{E} \left\{ (g(x_{1:N}) - \theta) \frac{\partial L(x_{1:N}|\theta)}{\partial \theta} \right\}^2
\end{aligned} \tag{2.41}$$

Finalmente, utilizando la desigualdad de Cauchy - Schwarz se obtiene

$$\begin{aligned}
1 &\leq \mathcal{E} \left\{ (\hat{\theta} - \theta)^2 \right\} \mathcal{E} \left\{ \left(\frac{\partial L(x_{1:N}|\theta)}{\partial \theta} \right)^2 \right\} \\
\mathcal{I}_F^{-1}(\theta) &= \frac{1}{\mathcal{E} \left\{ \left(\frac{\partial L(x_{1:N}, \theta)}{\partial \theta} \right)^2 \right\}} \leq \mathcal{E} \left\{ (\hat{\theta} - \theta)^2 \right\}
\end{aligned} \tag{2.42}$$

Esta demostración y su extensión (cuando $\Theta \in \mathbb{R}^{n_\theta}$) se puede apreciar con mayor detalle en [1, Apéndice 7A] y [2]. $\square$

Observación 2.9. El **Teorema 2.2** exige ciertas condiciones de regularidad (permiten introducir la derivada dentro de la integral), las cuales se encuentran detalladas en [1]. La demostración no requiere que $y_{1:n}$ sea una muestra aleatoria (no se utilizan sus propiedades), por lo cual el resultado es válido aun cuando las muestras no son independientes (ver [1, Capítulo 7]). ∇

Definición 2.5 (Eficiencia).

Sea $\hat{\theta}_N$ un estimador **insesgado** del parámetro θ basado en una muestra aleatoria (RS) de dimensión N . Entonces

- $\hat{\theta}$ es **eficiente** ssi su varianza es igual a la cota inferior de Cramér-Rao, es decir

$$\text{var}(\hat{\theta}_N) = \mathcal{E} \left\{ (\hat{\theta}_N - \theta) (\hat{\theta}_N - \theta)^T \right\} \equiv \mathcal{I}_F^{-1}(\theta) \quad (2.43)$$

- $\hat{\theta}$ es **asintóticamente eficiente** ssi su varianza es igual a la cota inferior de Cramér-Rao cuando $N \rightarrow \infty$, es decir

$$\lim_{N \rightarrow \infty} \text{var}(\hat{\theta}_N) = \lim_{N \rightarrow \infty} \mathcal{E} \left\{ (\hat{\theta}_N - \theta) (\hat{\theta}_N - \theta)^T \right\} \equiv \mathcal{I}_F^{-1}(\theta) \quad (2.44)$$

2.6. Lecturas complementarias

- En [1, Capítulo 14] y [2, Capítulo 5] es posible revisar los aspectos más relevantes de diseño del experimento y señales de entrada para obtener el conjunto de datos. En [88–90] es posible encontrar mayores detalles acerca del diseño del experimento.
- Para mayor información acerca de procesos estocásticos y variables aleatorias se sugiere revisar [83, 91–93]. Mientras que [18] se centra con mayor detalle en sistemas estocásticos, su estimación y control.
- Información adicional sobre conjuntos de modelos, sus parametrizaciones y aspectos relevantes se pueden revisar en [1, Capítulo 4 y 5], [2, Capítulo 6] y [3, Capítulo 4].
- En [72, 94] se encuentra un análisis de sistemas lineales en general.
- Diversos métodos de identificación, propiedades, aspectos relevantes tanto teóricos como prácticos se encuentran [1–3, 73, 77, 95, 96] y sus referencias, por mencionar algunos.
- La convergencia del estimador de máxima verosimilitud ML y del método del error de predicción PEM se puede revisar en [86, 97–99].

MÉTODO DE LOS MOMENTOS

El **método de los momentos** (MoM) es una técnica utilizada en la estimación paramétrica de modelos estadísticos. La idea básica consiste en encontrar valores para los parámetros de manera que los **momentos teóricos** (que implica el modelo) se aproximen tanto como sea posible a los **momentos muestrales** de los datos.

Esta metodología es atribuida a Pearson [100], quien la utilizó por primera vez para ajustar los parámetros de un modelo de mezcla sencillo.

Generalmente es considerado como un primo menor del **método de máxima verosimilitud** (ver **Sección 2.3**) debido a que este último posee propiedades teóricas superiores en diversas configuraciones. No obstante el método de los momentos y sus generalizaciones continúan siendo utilizadas en la práctica en ciertos problemas de estimación debido a su simpleza en términos conceptuales y computacionales.

A continuación se presenta el método de los momentos propuesto por Tugnait [16] para la identificación paramétrica de un sistema lineal invariante en el tiempo (LTI), de **fase no-mínima** (NMP) cuya entrada se encuentra dada por ruido blanco **no-Gaussiano** desconocido. La identificación se basa solamente en observaciones ruidosas de la salida del sistema.

Observación 3.1. *En caso de encontrarse familiarizado con el método de los momentos MoM, se sugiere continuar en el **Capítulo 4**, donde se aborda el problema de interés mediante máxima verosimilitud.*

3.1. Descripción del problema

Considere el siguiente sistema lineal e invariante en el tiempo

$$y_s(t) = - \sum_{i=1}^p a_i y_s(t-i) + \sum_{i=0}^{p-1} b_i w(t-i), \quad (3.1a)$$

con $b_0 = 1$,

$$y(t) = y_s(t) + v(t); \quad t \in \mathbb{N}, \quad (3.1b)$$

donde la señal $y_s(t)$ corresponde a un sistema ARMA($p, p-1$) (ver Tabla 2.1), la medición escalar contaminada con ruido se encuentra dada por $y(t)$, $\{w(t)\}$ representa una secuencia de ruido **no-gaussiana** escalar independiente e idénticamente distribuida (*i.i.d.*) con media cero y varianza σ_w^2 y $\{v(t)\}$ corresponde a una secuencia de ruido escalar, gaussiana *i.i.d.* con media igual a cero y varianza σ_v^2 . Las secuencias $v(t)$ y $w(t)$ son mutuamente independientes, es decir $w(t) \perp\!\!\!\perp v(t)$.

Entonces, a partir del vector de mediciones ruidosas $Y_{1:N} = \{y(t), 1 \leq t \leq N\}$ se desea estimar el vector de parámetros desconocidos θ , el cual se encuentra definido por

$$\theta = (a_1, a_2, \dots, a_p, b_1, b_2, \dots, b_{p-1}, \sigma_w^2, \sigma_v^2) \quad (3.2)$$

Dicho problema ha sido extensamente estudiado en la literatura de estimación y control, análisis de series de tiempo y procesamiento adaptativo de señales (ver e.g. [1, 3, 101, 102]). La mayoría de las técnicas de identificación existentes utilizan solo las estadísticas de segundo orden del conjunto de datos $Y_{1:N}$. Por ello, en el mejor de los casos, logran identificar el **sistema espectralmente equivalente de fase mínima** (SEMP) [37, 97, 103]. En otras palabras, un sistema de la forma (3.1) es identificable a partir del conjunto de datos $Y_{1:N}$ utilizando solo las estadísticas de segundo orden ssi el sistema es de fase mínima [38].

Se consideran conjuntos de datos extensos y procesos estacionarios en los siguientes desarrollos, por ello los sistemas que originan los datos deben ser exponencialmente estables (es decir que los polos del sistema deben encontrarse dentro del círculo unitario). Note que al obtenerse un sistema SEMP, los ceros también se encuentran restringidos al círculo unitario.

La motivación para considerar sistemas de fase no-mínima ha sido previamente considerada en [20, 37–40] entre muchos otros, por ende no será tratado aquí.

La identificación de sistemas de fase no-mínima ha sido abordado desde diferentes perspectivas. Por ejemplo en [37] la entrada es restringida a una secuencia de ruido Gaussiano, blanco, no estacionario con covarianza conocida. En [38] se limita el modelo a sistemas FIR sin ruido. Otro enfoque se puede apreciar en [20], donde la identificación se realiza al ajustar los espectros de alto orden (mayor que 2) entre las observaciones y el modelo.

Si bien existen diversas estrategias basadas en el métodos de los momentos para la identificación paramétrica de sistemas, nos centraremos en el desarrollo propuesto por Tugnait en [16].

3.2. Supuestos del Modelo

Sea $H[z]$ la función de transferencia (de tiempo discreto) del sistema definido por (3.1) con entrada $w(t)$ y salida $y_s(t)$. Entonces

$$H[z] = \sum_{i=0}^{p-1} b_i z^{-i} / \left[1 + \sum_{i=1}^p a_i z^{-i} \right]; \quad b_0 = 1; \quad (3.3)$$

Los siguientes supuestos son impuestos al modelo (3.1) (3.3)

Supuesto 3.1. Las raíces del polinomio denominador de la función de transferencia $H[z]$ (3.3) satisfacen $|z| < 1$. Es decir, el sistema es exponencialmente estable.

Supuesto 3.2. La secuencia i.i.d. no-Gaussiana de media cero $\{w(t)\}$ es tal que el cumulante de 4° orden γ_4 de la variable aleatoria $w(t)$ es distinto de cero

$$\gamma_4 = C_4(w(t)) \triangleq \mathcal{E} \{w^4(t)\} - 3 [\mathcal{E} \{w^2(t)\}]^2. \quad (3.4)$$

Supuesto 3.3. El ruido de medición $\{v(t)\}$ (3.1) es Gaussiano, i.i.d. con media igual a cero.

Supuesto 3.4. Sean p_0 y θ_0 los valores verdaderos del orden del sistema dado por (3.1) y del vector de parámetros a estimar (3.2) respectivamente. Adicionalmente se asume que la función de transferencia (3.3) no posee cancelaciones polo-cero con $p = p_0$ y $\theta = \theta_0$ y que $p_0 \geq 1$

Observación 3.2. *Variables no-Gaussianas con media cero y cumulante de 4° orden $\gamma_4 \neq 0$ incluyen la distribución de Bernouilli-Gaussiana considerada en [37], Uniforme, Chi-cuadrado, Maxwelliana [36, 106], entre muchas otras.* ∇

Observación 3.3. *El Supuesto 3.4 incluye modelos ARMA(p, q) donde $p < q$. Del mismo modo incluye modelos AR de dimensión finita (ver página 15)* ∇

Se define el vector aumentado de parámetros a estimar, de dimensión $(2p+2)$, como

$$\psi \triangleq (\theta, \gamma_4) \quad (3.5)$$

y el conjunto de parámetros factibles como

$$\Theta = \{\theta : (\text{Supuesto 3.1}) \text{ se cumple para } H(z|\theta)\} \quad (3.6)$$

3.3. Preliminares

En esta sección se revisan brevemente algunas definiciones y propiedades generales acerca de los cumulantes y momentos. Posteriormente se presentan definiciones necesarias para el desarrollo y se introduce la notación utilizada posteriormente en el capítulo.

3.3.1. Momentos y cumulantes

Los **momentos no centrales** μ'_k para una variable aleatoria x se encuentran definidos por

$$\mu'_k = \mathcal{E} \{x^n\}, \quad (3.7)$$

mientras que los **momentos centrales** μ_k se definen de la siguiente manera

$$\mu_k = \mathcal{E} \{(x - \mathcal{E} \{x\})^n\} \quad (3.8)$$

Los **cumulantes** $C_k(x)$ son estadísticas que proporcionan una alternativa a los momentos pese a encontrarse estrechamente relacionados.

Los momentos determinan los cumulantes en el sentido de que dadas dos distribuciones de probabilidad cuyos momentos sean idénticos, también tendrán cumulantes idénticos, y de manera similar, los cumulantes determinan los momentos.

El primer cumulante corresponde a la media, el segundo cumulante a la varianza y el tercer cumulante igual al tercer momento central. No obstante los cumulantes de cuarto orden y superiores no son iguales a los momentos centrales [107, 108]. En algunos problemas es preferible utilizar los cumulantes por sobre los momentos debido algunas propiedades atractivas de éstos. Dentro de las propiedades de los cumulantes destacan las siguientes [107]:

- Los cumulantes de tercer orden y superiores de una distribución gaussiana son cero, siendo ésta la única distribución que presenta esta característica.
- Dadas dos variables aleatorias x e y independientes, es decir $x \perp\!\!\!\perp y$, se tiene que el cumulante de orden n de la suma de ellas corresponde a la suma de los correspondientes cumulantes de los sumandos

$$C_n(x + y) = C_n(x) + C_n(y), \quad \text{si } x \perp\!\!\!\perp y. \quad (3.9)$$

- Considere una variable aleatoria x y una constante c , entonces se tiene que el primer cumulante es equivariante y los demás son invariantes, es decir

$$C_1(x + c) = C_1(x) + c \quad (3.10a)$$

$$C_n(x + c) = C_n(x), \quad \forall n \geq 2 \quad (3.10b)$$

en otras palabras, al desplazar una variable aleatoria (agregando una constante c) solo modifica el primer cumulante y no afecta a los demás.

- El cumulante de orden n es homogéneo de grado n , o sea

$$C_n(cx) = c^n C_n(x), \quad (3.11)$$

donde x representa una variable aleatoria y c una constante.

Estas y otras propiedades de los cumulantes se pueden revisar en [107, 108].

A continuación se presentan algunas relaciones entre los cumulantes C_n y los momentos centrales μ_n (se ha omitido la variable aleatoria (x) en las siguientes expresiones)

$$C_2 = \mu_2 \quad (3.12a)$$

$$C_3 = \mu_3 \quad (3.12b)$$

$$C_4 = \mu_4 - 3\mu_2^2 \quad (3.12c)$$

$$C_5 = \mu_5 - 10\mu_3\mu_2 \quad (3.12d)$$

$$C_6 = \mu_6 - 15\mu_4\mu_2 - 10\mu_3^2 + 30\mu_2^3 \quad (3.12e)$$

$$C_7 = \mu_7 - 21\mu_5\mu_2 - 35\mu_4\mu_3 + 210\mu_3\mu_2^2 \quad (3.12f)$$

Adicionalmente se tiene que el primer cumulante es igual a la media $C_1 = \mu'_1$. Las relaciones entre los momentos no centrales μ'_k , momentos centrales μ_k y cumulantes C_k se pueden revisar en detalle en [109].

A partir de la ley de los grandes números es posible obtener estimadores para los momentos (ver (3.7) y (3.8)), los cuales se denominan **momentos muestrales**. Luego, a partir de las relaciones entre cumulantes y momentos (3.12) se obtienen los cumulantes muestrales. A continuación se pueden apreciar los cumulantes muestrales hasta orden 4

$$\hat{C}_1 = \frac{1}{N} \sum_{i=1}^N x_i \quad (3.13a)$$

$$\hat{C}_2 = \frac{1}{N} \sum_{i=1}^N (x_i - \hat{C}_1)^2 \quad (3.13b)$$

$$\hat{C}_3 = \frac{1}{N} \sum_{i=1}^N (x_i - \hat{C}_1)^3 \quad (3.13c)$$

$$\hat{C}_4 = \left(\frac{1}{N} \sum_{i=1}^N (x_i - \hat{C}_1)^4 \right) - 3\hat{C}_2^2, \quad (3.13d)$$

donde N corresponde al número de muestras. Cabe destacar que los cumulantes muestrales obtenidos de esta manera son **asíntoticamente insesgados y consistentes** [109].

Las varianzas del error de estimación asociadas a los cumulantes muestrales (3.13) se encuentran dadas por

$$\text{var}(C_1) = P_{C_1} = \frac{1}{N}\mu_2 \quad (3.14a)$$

$$\text{var}(C_2) = P_{C_2} = \frac{1}{N}(\mu_4 - \mu_2^2) \quad (3.14b)$$

$$\text{var}(C_3) = P_{C_3} = \frac{1}{N}(\mu_6 - \mu_3^2 + 9\mu_2^3 - 6\mu_2\mu_4) \quad (3.14c)$$

$$\text{var}(C_4) = P_{C_4} = \frac{1}{N}(\mu_8 - 12\mu_6\mu_2 - 8\mu_5\mu_3 - \mu_4^2 + 48\mu_4\mu_2^2 + 64\mu_3^2\mu_2 - 36\mu_2^4), \quad (3.14d)$$

donde N corresponde al número de muestras. En [24] se presentan dichos resultados y se analiza el *efecto de un número limitado de muestras N* en los cumulantes muestrales. En [109] es posible revisar estimadores insesgados y consistentes para los cumulantes, los cuales se denominan **estadísticas- k** , junto a sus varianzas.

Con el fin de ilustrar el efecto del número de datos en los cumulantes muestrales se presentan a continuación los resultados de una simulación numérica. Considere que se cuentan con n realizaciones de la variable aleatoria x , la cual posee una distribución *Chi-cuadrado*, i.e. $x \sim \chi^2(\kappa)$, con $\kappa = 2$. Los cumulantes (teóricos) para dicha distribución son conocidos y se encuentran dados por

$$C_k(x) = 2^{k-1}(k-1)!\kappa, \quad \forall x \sim \chi^2(\kappa) \quad (3.15)$$

En la Fig. 3.1 se pueden observar los cumulantes teóricos $C_k(x)$ de primer a cuarto orden obtenidos a partir de (3.15) y los cumulantes muestrales $\hat{C}_k(x)$ obtenidos a partir de (3.13) para todo n .

Note que la cantidad de muestras requeridas para que los cumulantes muestrales convergan aumenta junto al orden de estos (ver las diferencias en las escalas). En la Fig. 3.2 se presentan los resultados obtenidos a partir de la simulación numérica para la varianza teórica de los cumulantes muestrales $P_{C_k(x)}$ obtenidos mediante (3.14) y las varianzas muestrales $\hat{P}_{C_k(x)}$ obtenidas a partir de la simulación numérica.

En la Fig. 3.2 se puede apreciar que el error o incertidumbre asociado a los cumulantes muestrales aumenta junto con el orden.

Entonces, a partir de los resultados obtenidos es posible afirmar que los *cumulantes muestrales* (y debido a su estrecha relación, los *momentos muestrales*) de alto orden requieren un número de muestras elevado para producir estimaciones precisas.

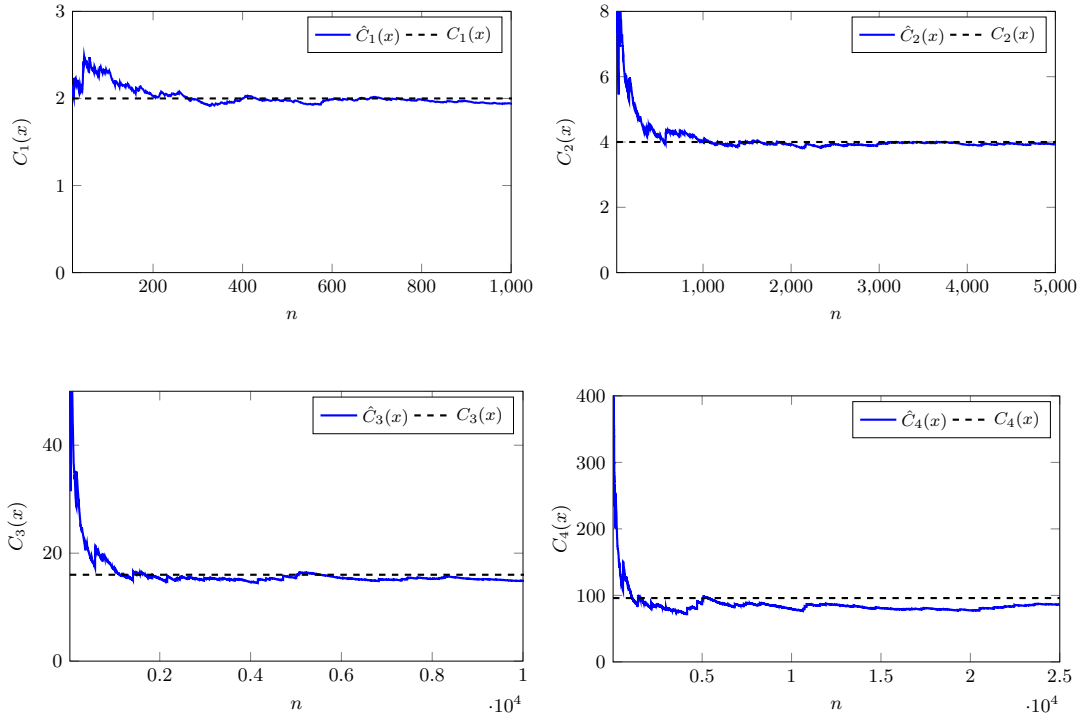


Fig. 3.1: Cumulantes muestrales $\hat{C}_k(x)$ versus cumulantes teóricos $C_k(x)$ a partir de n realizaciones.

3.3.2. Cumulantes en procesos aleatorios y sistemas

A continuación se presentan algunas definiciones y propiedades de los cumulantes para procesos aleatorios.

Para un vector de datos $Y_{1:N} = \{y(t_1), y(t_2), \dots, y(t_k)\}$ con $\mathcal{E}\{y(t_i)^k\} < \infty$ $i = 1, 2, \dots, k$, denotaremos como $\text{Cum}_k(y(t_1), y(t_2), \dots, y(t_k))$ al k -ésimo cumulante de Y^N . Para una serie de tiempo estacionaria $\{y(t_1), y(t_2), \dots, y(t_k)\}$ y para todo retardo de tiempo τ se tiene que

$$\begin{aligned} \text{Cum}_k(y(t), y(t+t_1), \dots, y(t+t_{k-1})) &= \text{Cum}_k(y(t+\tau), y(t+\tau+t_1), \dots, y(t+\tau+t_{k-1})) \\ &\triangleq C_k(t_1, t_2, \dots, t_{k-1}) \end{aligned} \quad (3.16)$$

Para una serie estacionaria de media cero $\{y(t)\}$ se puede demostrar que [110, 111]

$$\begin{aligned} C_4(t_1, t_2, t_3) &= R_4(t_1, t_2, t_3) - R_2(t_1)R_2(t_3 - t_2) \dots \\ &\quad - R_2(t_2)R_2(t_3 - t_1) - R_2(t_3)R_2(t_2 - t_1), \end{aligned} \quad (3.17)$$

donde

$$R_4(t_1, t_2, t_3) \triangleq \mathcal{E}\{y(t)y(t+t_1)y(t+t_2)y(t+t_3)\} \quad (3.18)$$

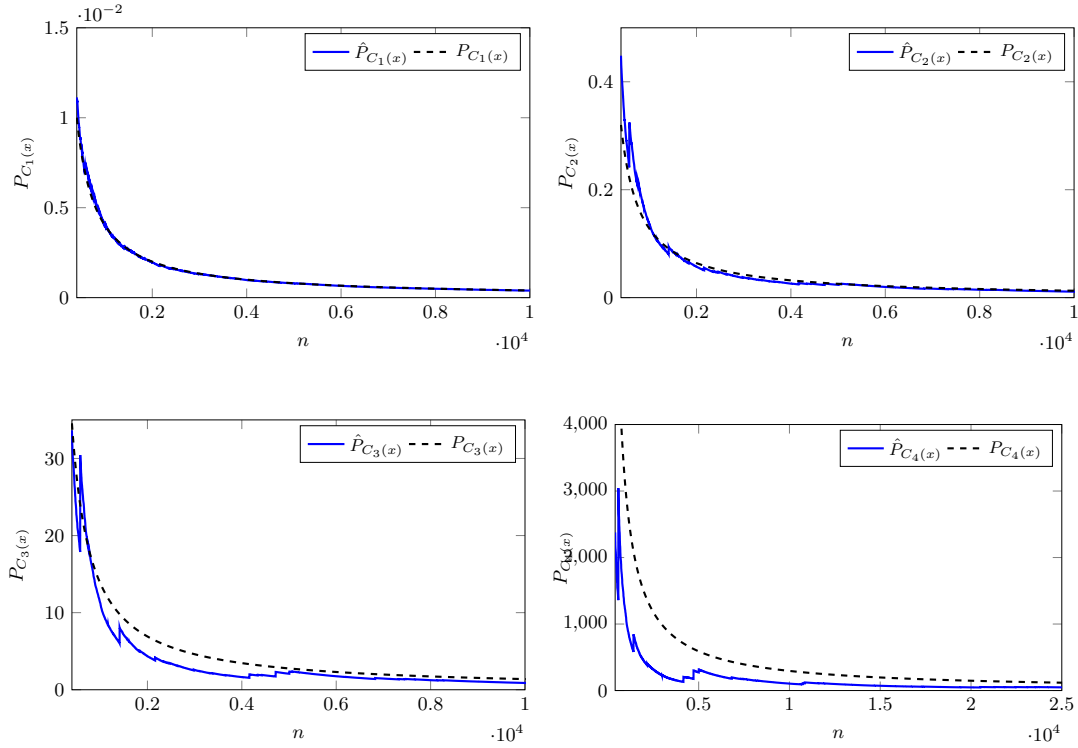


Fig. 3.2: Varianza teórica $P_{C_k(x)}$ y muestral $\hat{P}_{C_k(x)}$ de los cumulantes muestrales a partir de n realizaciones.

$$R_2(\tau) \triangleq \mathcal{E} \{y(t)y(t+\tau)\} \quad (3.19)$$

Adicionalmente, para una variable aleatoria gaussiana, el cumulante de orden k -ésimo γ_k desaparece para $k > 2$. Del mismo modo, para un proceso gaussiano el cumulante de orden k -ésimo γ_k desaparece para $k > 2$. Por ende, dado que $v(t)$ en (3.1) distribuye de manera Normal, se tiene que

$$C_4(t_1, t_2, t_3) = C_{s4}(t_1, t_2, t_3) = \text{Cum}(y_s(t)y_s(t+t_1)y_s(t+t_2)y_s(t+t_3)) \quad (3.20)$$

Al aplicar la transformada Zeta bilateral al cumulante de 4° orden $C_4(t_1, t_2, t_3)$ se obtiene

$$S_4(z_1, z_2, z_3) = \sum_{t_1=-\infty}^{\infty} \sum_{t_2=-\infty}^{\infty} \sum_{t_3=-\infty}^{\infty} C_4(t_1, t_2, t_3) z_1^{-t_1} z_2^{-t_2} z_3^{-t_3} \quad (3.21)$$

Para un modelo dado por (3.1) en el cual se cumplan los supuestos presentados en la **Sección 3.2** y (3.20) se puede probar que [110–112]

$$S_4(z_1, z_2, z_3) = \gamma_4 H(z_1)H(z_2)H(z_3)H([z_1 z_2 z_3]^{-1}). \quad (3.22)$$

Al reemplazar z_i por $e^{j\omega_i}$ con $i = (1, 2, 3)$ y $j = \sqrt{-1}$ se obtiene $S_4(e^{j\omega_1}, e^{j\omega_2}, e^{j\omega_3})$ el cual se denomina **Triespectro** de la serie de tiempo $\{y(t)\}$.

En [20] se demuestra que para estimar la fase de un sistema es necesario el uso de espectros de alto orden (o de manera equivalente, momentos o cumulantes de alto orden) junto a la densidad espectral de potencia (PSD).

Lema 3.1. *Considere la salida $y_s(t)$ de un sistema definido por (3.1). Adicionalmente asuma que $H(1) \neq 0$ y que los **Supuestos 3.1 y 3.4** definidos en la **Sección 3.2** se cumplen. Entonces, la función de transferencia $H(z)$ se puede determinar para $z = \exp(j\omega)$, $-\pi \leq \omega < \pi$ a partir de la **densidad espectral de potencia (PSD)** y el triespectro de $\{y_s(t)\}$.*

Demostración 3.1. Ver [113, Página 349-350]. □

A continuación se resumen algunos resultados acerca de la obtención del cumulante de 4° orden del proceso $\{y(t)\}$ dado por (3.1). La dependencia de los parámetros y variables contenidos en θ y ψ serán explicitados en las siguientes ecuaciones. Utilizando las propiedades presentes en [111, Página 211] y [110, Página 1357] se puede demostrar que

$$C_4(t_1, t_2, t_3 | \psi) = \gamma_4(\psi) \sum_{i=0}^{\infty} h(t_1 + i, \theta) h(t_2 + i, \theta) h(t_3 + i, \theta) h(i, \theta), \quad (3.23)$$

donde $h(i, \theta)$, $i \geq 0$ representa la respuesta a impulso del modelo parametrizado por θ , es decir satisface la siguiente ecuación recursiva

$$h(i, \theta) = - \sum_{j=1}^p a_j(\theta) h(i-j, \theta) + \sum_{j=0}^{p-1} b_j(\theta) u(i-j), \quad (3.24)$$

con $u(i) = \delta(i)$ y $h(i, \theta) = 0$ para $i < 0$. Este resultado se puede demostrar al calcular los momentos de un sistema de media móvil (MA) o respuesta a impulso infinita (IIR) de manera análoga al presentado en [2, 18] para el cálculo de las funciones de covarianza. Debido al **Supuesto 3.1** se sabe que $h(i, \theta) \approx 0$ para algún entero positivo $i \geq F$. Por consiguiente, en la practica, es posible reemplazar el límite superior de (3.23) por F . Note que F puede ser una función de θ (ver e.g. **tiempo de asentamiento** en [76, 114]).

En [115] se presenta una manera alternativa de calcular los cumulantes teóricos a partir de una representación en variables de estado del sistema utilizando el **producto de Kronecker**. Otros autores han utilizado las ecuaciones de **Yule-Walker** y su versión modificada para el calculo de los cumulantes teóricos principalmente para sistemas AR y en algunos casos ARMA (ver e.g. [116]).

Observación 3.4. *De manera análoga a la función de covarianza (segundo momento) existen ciertas simetrías en los cumulantes. En el caso del cumulante de 4° orden se*

tiene que

$$\begin{aligned}
 C_4(t_1, t_2, t_3|\psi) &= C_4(-t_3, t_1 - t_3, t_2 - t_3|\psi) \\
 &= C_4(-t_2, t_1 - t_2, t_3 - t_2|\psi) \\
 &= C_4(-t_1, t_2 - t_1, t_3 - t_1|\psi)
 \end{aligned} \tag{3.25}$$

con $-\infty < t_i < \infty$, $i = 1, 2, 3$ (ver [110–112]). ∇

3.4. Procedimiento de Identificación

A continuación se presenta el método de identificación propuesto por Tugnait [16], el cual consiste de 2 pasos.

3.4.1. Estimación de un modelo espectralmente equivalente

En el primer paso del procedimiento se puede utilizar cualquier método que obtenga una estimación consistente (ver **Definición 2.4**) de un modelo espectralmente equivalente.

Por simplicidad utilizaremos el **método del error de predicción** (PEM), el cual fue abordado previamente en el **Capítulo 2** (particularmente en la **Sección 2.4**), no obstante se presentan a continuación algunos pasos cruciales para facilitar la discusión del segundo paso. (para mayor detalle u otro método de identificación que obtenga una estimación consistente de un modelo espectralmente equivalente se sugiere revisar [1–3, 102]).

En el método del error de predicción se minimiza (Nota: En algunos casos se omite la varianza) el siguiente funcional de costo escalar respecto de θ sujeto al modelo (3.1) y el **Supuesto 3.1**.

$$J_N(\theta) = \frac{1}{2} \sum_{i=1}^N [\tilde{y}^2(i|\theta)/P(i|\theta) + \log P(i|\theta)], \tag{3.26}$$

donde $\tilde{y}(i|\theta) = (y(i) - \hat{y}(i|\theta))$ representa el error de predicción, $\hat{y}(i|\theta)$ corresponde al predictor a un paso para $y(i)$ dada la colección $\{y(t)\}_{1:N-1}$. $P(i|\theta)$ representa la varianza del error de predicción, definida por

$$P(i|\theta) = \mathcal{E} \left\{ (y(i) - \hat{y}(i|\theta))^2 \right\} \tag{3.27}$$

El funcional de costo, dado por (3.26), se minimiza mediante un método de optimización convexo, usualmente basado en gradiente como el de Gauss-Newton [1, 84, 117].

Sea $\hat{\theta}_N$ el parámetro estimado a partir de los datos $\{y(t)\}_{1:N}$ obtenido al minimizar (3.26) (u obtenida mediante algún método consistente basado en estadísticas de 2° orden). El modelo obtenido corresponde a un sistema espectralmente equivalente, cuyos polos se encuentran dentro del círculo unitario. La función de transferencia del modelo (ignorando el ruido de medición) se encuentra dada por

$$\hat{H}(z) = H(z|\hat{\theta}_N) = \frac{z^{p-1} + \sum_{i=1}^{p-1} \hat{b}_i z^{p-i}}{z^p + \sum_{i=1}^p \hat{a}_i z^{p-i}}, \quad (3.28)$$

con $\hat{b}_i = b_i(\hat{\theta}_N)$ y $\hat{a}_i = a_i(\hat{\theta}_N)$. El polinomio numerador de (3.28) posee $p - 1$ raíces. Suponga que posee n_r raíces reales y n_c raíces complejas de modo que

$$n_r + n_c = p - 1 \quad (3.29)$$

Dado que los coeficientes de $\hat{b}_i$ son reales, entonces las raíces complejas deben encontrarse en pares conjugados, por ende n_c es un entero par. **La densidad espectral de potencia** (PSD) [1, 18] del modelo (omitendo el ruido de medición) se encuentra dado por

$$S_2(z) = \sigma_w^2(\hat{\theta}_N) H(z|\hat{\theta}_N) H(z^{-1}|\hat{\theta}_N) \quad (3.30)$$

con $z = \exp(j\omega)$. Consecuentemente existen (a lo más) $2^{n_r+n_c/2}$ combinaciones posibles de ceros que poseen, salvo por una constante de escalamiento, la misma densidad espectral de potencia (3.30). Un cero ubicado en f o en $1/f$ producen el mismo producto salvo por una constante de escalamiento

$$(z - f)(z^{-1} - f) = f^2(z - f^{-1})(z^{-1} - f^{-1}) \quad (3.31)$$

Observación 3.5. Si el sistema posee ceros ubicados sobre el círculo unitario, entonces el número de combinaciones posibles de ceros que conllevan la misma densidad espectral de potencia será menor a $2^{n_r+n_c/2}$. ∇

Sea Θ_N el conjunto de parámetros de posibles combinaciones del vector de parámetros $\hat{\theta}_N$ considerando las posibles ubicaciones de los ceros para el **modelo espectralmente equivalente**. Note que los elementos de Θ_N difieren en los parámetros b_i y σ_w^2 , dado que al modificar los coeficientes b_i es necesario modificar σ_w^2 para mantener una densidad espectral de potencia idéntica (ver (3.31)).

Sea α_N un elemento arbitrario contenido en Θ_N . Note que todos los elementos α_N poseen el mismo conjunto de polos y varianza del ruido σ_v^2 . Entonces resumiendo, al

final del primer paso se obtiene un conjunto de modelos Θ_N que contiene a lo más 2^{p-1} elementos (si todos los ceros son reales). Dicho conjunto se puede definir como

$$\Theta_N = \left\{ \theta : S_{2y}(z|\theta) = S_{2y}(z|\hat{\theta}_N) \right\}, \quad (3.32)$$

donde

$$S_{2y}(z|\theta) = \sigma_w^2(\theta)H(z|\theta)H(z^{-1}|\theta) + \sigma_v^2(\theta). \quad (3.33)$$

3.4.2. Determinación de la ubicación correcta de los ceros

En el segundo paso se busca estimar la posición correcta de los ceros del sistema (3.1) a partir del conjunto de datos $Y_{1:N}$ y el conjunto de modelos Θ_N (3.32).

En otras palabras, se selecciona un elemento α_N a partir de Θ_N de modo que posea el menor error cuadrático medio entre la función de cumulantes teóricos y los cumulantes muestrales obtenidos a partir del conjunto de datos $Y_{1:N}$.

El cumulante muestral de 4° orden para el conjunto de datos $Y_{1:N}$ se estima de la siguiente manera

$$\begin{aligned} \hat{C}_4(t_1, t_2, t_3) = & \frac{1}{N} \left[\sum_{t=0}^{N-t_M} y(t)y(t+t_1)y(t+t_2)y(t+t_3) \right] - \hat{R}_2(t_1)\hat{R}_2(t_3-t_2)... \\ & - \hat{R}_2(t_2)\hat{R}_2(t_3-t_1) - \hat{R}_2(t_3)\hat{R}_2(t_2-t_1), \end{aligned} \quad (3.34)$$

con $t_M = \max(t_1, t_2, t_3)$ y

$$\hat{R}_2(i) = \frac{1}{N} \sum_{t=0}^{N-i} y(t)y(t-i). \quad (3.35)$$

Las incógnitas son γ_4 y $\alpha_N \in \Theta_N$, las cuales se seleccionan de manera tal que minimicen el criterio cuadrático dado por

$$J_{2N} = \sum_{t_1=0}^L \sum_{t_2=0}^L \sum_{t_3=0}^L \left[C_4(t_1, t_2, t_3|\alpha_N, \gamma_4) - \hat{C}_4(t_1, t_2, t_3) \right]^2, \quad (3.36)$$

con C_4 y $\hat{C}_4$ definidos en (3.23) y (3.34) respectivamente y para un L suficientemente grande. En [16, Sección 5] se demuestra que para $L \geq 2p$ la estimación paramétrica es consistente (ver **Definición 2.4**).

Para estimar γ_4 se deriva el funcional de costo dado por (3.36) con respecto a γ_4 para un α_N fijo. Luego se iguala dicha derivada a cero para obtener el siguiente resultado

$$\hat{\gamma}_4(\alpha_N) = \frac{\sum_{t_1=0}^L \sum_{t_2=0}^L \sum_{t_3=0}^L \hat{C}_4(t_1, t_2, t_3) \bar{C}_4(t_1, t_2, t_3|\alpha_N)}{\sum_{t_1=0}^L \sum_{t_2=0}^L \sum_{t_3=0}^L [\bar{C}_4(t_1, t_2, t_3|\alpha_N)]^2}, \quad (3.37)$$

donde $\hat{C}_4$ se encuentra dado por (3.34) y $\bar{C}_4$ se obtiene mediante

$$\bar{C}_4 = \frac{C_4(t_1, t_2, t_3 | \alpha_N, \gamma_4)}{\gamma_4} \quad (3.38)$$

con C_4 obtenido mediante (3.23). Sustituya (3.37) en (3.36) y elija $\alpha_N \in \Theta_N$ que minimice el costo J_{2N} .

Observación 3.6. Si bien el caso presentado corresponde a un modelo ARMA (3.1), es posible extender fácilmente el procedimiento de identificación a sistemas con entrada exógena. Para ello obtenga en primer lugar un modelo espectralmente equivalente con PEM (ver **Subsección 3.4.1**) y posteriormente determine la ubicación correcta de los ceros a partir de del sistema auxiliar $y_{aux} = y_t - G(z|\hat{\theta}_N)u_t$ utilizando el procedimiento descrito en la **Subsección 3.4.2**. ∇

A continuación se presentan algunos resultados que conciernen el comportamiento del estimador previamente mencionado para un numero de muestras N elevado.

Lema 3.2. Bajo el **Supuesto 3.1** se tiene que para todo conjunto de enteros $(t_1, t_2, t_3) \in \mathbb{N}$ el cumulante muestral $\hat{C}_4$ converge con probabilidad uno al cumulante de 4° orden, es decir

$$\lim_{N \rightarrow \infty} \hat{C}_4(t_1, t_2, t_3) \stackrel{w.p.1.}{=} Cum_4(y(t), y(t+t_1), y(t+t_2), y(t+t_3)) = C_{s4}(t_1, t_2, t_3 | \theta_0, \gamma_{40}). \quad (3.39)$$

Demostración 3.2. Ver [92, 118]. $\square$

Lema 3.3. Considere la minimización del funcional de costo dado por (3.26) respecto de θ sujeto al sistema (3.1) y Θ definido en (3.6). Adicionalmente asuma que $p \geq p_0$ y $\theta_0 \in \Theta$, es decir el sistema no se encuentra submodelado. Entonces el parámetro $\hat{\theta}$ estimado mediante el **método del error de predicción** converge con probabilidad uno al conjunto D_1 definido por

$$D_1 = \{\theta : \theta \in \Theta, S_{2y}(z|\theta) = S_{2y}(z|\theta_0)\} \subseteq \Theta \quad (3.40)$$

Es decir, converge a un modelo SE.

Demostración 3.3. Ver [1, 86, 98, 119]. $\square$

Lema 3.4. Bajo los supuestos del **Lema 3.3** y el **Supuesto 3.4** se tiene que

$$\sigma_v^2(\theta) = \sigma_v^2(\theta_0), \quad \forall \theta \in D_1. \quad (3.41)$$

Demostración 3.4. *Evalúe la igualdad dada por (3.40) para $z = 0$. Debido al **Supuesto 3.4** y (3.33) se tiene que el producto $H(z|\theta)H(z^{-1}|\theta)$ desaparece cuando $z = 0$ para todo $\theta \in D_1$. Para mayores detalles ver [16, 101].* $\square$

Lema 3.5. *Considere los supuestos del **Lema 3.4** y suponga adicionalmente que $p = p_0$, $\gamma_{40} \neq 0$ y que no existen cancelaciones polo-cero en el sistema real $H(z|\theta_0)$ ni en su equivalente espectral de fase mínima obtenido al reflejar los ceros ubicados en el exterior del círculo unitario hacia su interior. Entonces, el estimador paramétrico definido en la **Sección 3.4** es fuertemente consistente (también denominada convergencia casi segura).*

Demostración 3.5. *Ver [16, Apéndice].* $\square$

El problema de determinación del orden del modelo p no sera abordado en profundidad en este trabajo, dado que se asume conocido. Además estamos interesados principalmente en el análisis del desempeño de los estimadores. De cualquier modo, la determinación del orden debe considerarse dentro del primer paso del procedimiento propuesto en la **Sección 3.4**. Un criterio ampliamente utilizado para este fin (no obstante existen varios en la literatura, ver e.g. [2, 120]) es el **criterio de información de Akaike** [121]. No obstante trabajos posteriores demostraron que dicho criterio es inconsistente [120, 122]. En [120, 122] se propone una modificación al criterio de Akaike, el cual se conoce como **MAIC**, para estimar el orden del sistema y se encuentra dado por

$$\text{MAIC}(p) = 2J_N(\hat{\theta}) + (2p + 1) \log(N) + N \log(2\pi) \quad (3.42)$$

donde $J_N(\theta)$ corresponde al funcional de costo dado por (3.26), $\hat{\theta}$ corresponde al argumento que minimiza $J_N(\theta)$ dado p y el número de parámetros desconocidos se encuentra dado por $2p + 1$. Para seleccionar el orden del modelo elija un límite superior para el orden del modelo, dado por $\bar{p}$ y calcule el criterio modificado de Akaike $\text{MAIC}(p)$ para $1 \leq p \leq \bar{p}$. Finalmente seleccione el valor de p que minimice $\text{MAIC}(p)$.

3.5. Lecturas complementarias

- En [107, 108] se encuentran las definiciones, propiedades y varios resultados asociados a los cumulantes y momentos. El efecto del número de muestras en los cumulantes muestrales se puede revisar en [24].
- En [95, Página 351-355] se presenta una versión resumida y básica del método de los momentos para la estimación de sistemas lineales, junto con ejemplos de MATLAB.

- Otras estrategias utilizadas en la identificación paramétrica de sistemas ARMA se pueden revisar en [21, 56, 123, 124, 124] y las referencias presentes en ellos.
- En [125, Capítulo 3] se presentan diversas estrategias del método de estimación paramétrica basada en estadísticas de alto orden para modelos AR, MA y ARMA. Además se presentan algunos ejemplos, aplicaciones, simulaciones y una estimación insesgada de los momentos y cumulantes basadas en las estadísticas- k [126, 127].
- En [128] se encuentra un análisis del desempeño de diversos algoritmos para la identificación paramétrica de sistemas basados en estadísticas de alto orden.
- Un análisis de la identificabilidad (supuestos, convergencia) de modelos ARMA no-Gaussianos utilizando cumulantes se presenta en [17].
- Un compendio del uso de estadísticas de alto orden aplicadas a diversos usos de procesamiento de señales, dentro de las cuales se incluye estimación paramétrica de sistemas, se encuentra en [129].
- En [130] se puede apreciar una biografía de las estadísticas de alto orden junto a una clasificación de sus usos y diversas aplicaciones.
- MATLAB posee un toolbox de análisis espectral de alto orden, que incluye el método de los momentos para la estimación paramétrica de sistemas ARMA [131].¹

¹Toolbox disponible en <https://www.mathworks.com/matlabcentral/fileexchange/3013-hosa-higher-order-spectral-analysis-toolbox>
Documentación disponible en <https://labcit.ligo.caltech.edu/~rana/mat/HOSA/HOSA.PDF>

IDENTIFICACIÓN MEDIANTE ML DE SISTEMAS CON RUIDO GMM

Este capítulo se centra en el desarrollo y análisis de una solución al problema de interés definido previamente en la **Sección 1.1**. Para ello se considera en primer lugar el problema de estimación de una constante en presencia de ruido aditivo **no-gaussiano**, cuyo objetivo es ilustrar el efecto sobre la precisión del estimador, dada por la covarianza del error de estimación, cuando se conoce o aproxima la distribución del ruido.

Posteriormente, se analiza la función de verosimilitud obtenida en el problema previo.

Luego, se desarrolla la expresión que define el problema de estimación de máxima verosimilitud para sistemas dinámicos, considerando que la distribución del ruido se puede aproximar, o corresponde, a un modelo de mezcla de distribuciones gaussianas GMM. Dado que el análisis de la función de verosimilitud muestra que el problema es **multi-modal**, y por ende **no-convexo**, es necesario utilizar un algoritmo de optimización global para resolver el problema de estimación.

Finalmente se analiza el desempeño del método propuesto mediante ejemplos numéricos. Los resultados son contrastados con el método del error de predicción PEM.

Cabe destacar que en este capítulo se aborda solamente el problema de **fase mínima** para la función de transferencia asociada al ruido. La extensión del método aquí propuesto se encuentra en el **Capítulo 5**.

4.1. Estimación de una constante

Con el fin de ilustrar la ventaja de conocer o aproximar la distribución del ruido al utilizar el **método de máxima verosimilitud** ML (ver **Sección 2.3**) se analiza a continuación el problema más simple de estimación, el cual se encuentra dado por mediciones de una constante en presencia de ruido aditivo.

Se asume que la distribución del ruido se encuentra dado por una mezcla de distribucio-

nes Gaussianas (4.2), también denominado **Modelo de Mezcla Gaussiana** (GMM). Considere entonces que el *sistema verdadero* a partir del cual se generan los datos, se encuentra dado por

$$y_t = \theta_0 + \omega_t, \quad (4.1)$$

donde $\theta_0 \in \mathbb{R}$ corresponde al parámetro escalar a ser estimado, y_t a la salida del sistema y ω_t a ruido blanco aditivo de media cero $\mu_\omega = 0$.

La distribución del ruido ω_t se encuentra dada por el siguiente GMM

$$p(\omega_t) = \sum_{j=1}^M \alpha_j \phi(\omega_t; \mu_j, \sigma_j^2) \quad \text{sueto a} \quad \sum_{j=1}^M \alpha_j = 1, \quad (4.2)$$

con $M \geq 2$ ($M \in \mathbb{N}$), α_j corresponde al j -ésimo peso de la mezcla y $\phi(\omega_t; \mu_j, \sigma_j^2)$ representa una distribución gaussiana o Normal con media μ_j y varianza σ_j^2 dada por

$$\phi(\omega_t; \mu_j, \sigma_j^2) = \frac{1}{\sqrt{2\pi}\sigma_j} \exp \left\{ -\frac{(\omega_t - \mu_j)^2}{2\sigma_j^2} \right\} \quad (4.3)$$

A continuación se analiza la precisión, en términos de varianza del error de estimación, de los estimadores de **mínimos cuadrados** (LS) y **máxima verosimilitud** (ML) aplicados al sistema dado por (4.1).

Con la finalidad de simplificar la presentación y sin pérdida de generalidad se analiza el caso cuando $M = 2$.

4.1.1. Estimación mediante mínimos cuadrados (LS)

Es sabido que el estimador de mínimos cuadrados LS¹ para θ_0 en (4.1) se encuentra dado por

$$\hat{\theta}_{LS} = \frac{1}{N} \sum_{t=1}^N y_t. \quad (4.4)$$

Si bien el siguiente resultado es ampliamente conocido, es incluido por razones de completitud.

Lema 4.1. *La varianza del error de estimación del estimador de cuadrados mínimos, $cov\{\hat{\theta}_{LS}\}$, para el sistema (4.1) con $M = 2$ se encuentra dado por:*

$$cov\{\hat{\theta}_{LS}\} = \frac{1}{N} \sum_{j=1}^2 \alpha_j (\sigma_j^2 + \mu_j^2) \quad (4.5)$$

¹El método de mínimos cuadrados (LS) es equivalente al método del error de predicción (PEM) para el sistema dado por (4.1) (Ver e.g. [1, 2]).

Demostración 4.1. *El estimador LS para θ_0 en (4.1) se encuentra dado por:*

$$\hat{\theta}_{LS} = \frac{1}{N} \sum_{t=1}^N y_t \stackrel{(a)}{=} \theta_0 + \frac{1}{N} \sum_{t=1}^N \omega_t \quad (4.6)$$

Donde (a) se obtiene al reemplazar (4.1) en (4.6). Dado que la media del ruido ω_t es cero, se tiene que el estimador LS es **insesgado** (ver Definición 2.3) y que converge de manera casi segura (a.s.) a θ_0 , es decir $\hat{\theta}_{LS} \xrightarrow{a.s.} \theta_0$, puesto que

$$\mathcal{E}\{\hat{\theta}_{LS}\} = \theta_0 + \frac{1}{N} \sum_{t=1}^N \mathcal{E}\{\omega_t\} \stackrel{(a)}{=} \theta_0, \quad (4.7)$$

donde $\mathcal{E}\{\cdot\}$ corresponde al operador esperanza y (a) es consecuencia de que $\mu_j = 0 \forall j$. La covarianza del estimador LS se obtiene de la siguiente manera:

$$\text{cov}\{\hat{\theta}_{LS}\} = \mathcal{E} \left\{ \left[\frac{1}{N} \sum_{t=1}^N \omega_t \right]^2 \right\} = \frac{1}{N^2} \sum_{t=1}^N \mathcal{E}\{\omega_t^2\}. \quad (4.8)$$

Utilizando la definición del operador esperanza y la función de densidad de probabilidad dada por (4.2) se tiene

$$\begin{aligned} \mathcal{E}\{\omega_t^2\} &= \int_{-\infty}^{\infty} \omega_t^2 p(\omega_t) d\omega_t \\ &= \sum_{j=1}^K \alpha_j \left[\int_{-\infty}^{\infty} (\omega_t - \mu_j + \mu_j)^2 \phi(\omega_t; \mu_j, \sigma_j^2) d\omega_t \right] \\ &= \sum_{j=1}^K \alpha_j (\sigma_j^2 + \mu_j^2) \end{aligned} \quad (4.9)$$

Finalmente al reemplazar (4.9) en (4.8) y con $K = 2$, se obtiene la covarianza del estimador LS:

$$\text{cov}\{\hat{\theta}_{LS}\} = \frac{1}{N} \sum_{j=1}^2 \alpha_j (\sigma_j^2 + \mu_j^2) \quad (4.10)$$

□

4.1.2. Estimación mediante máxima verosimilitud (ML)

El estimador de máxima verosimilitud (ML), $\hat{\theta}_{ML}$, converge de manera asintótica en distribución como $\sqrt{N} (\hat{\theta}_{ML} - \theta_0) \xrightarrow{d} \mathcal{N}(0, \mathcal{I}_F^{-1}(\theta))$, donde $\mathcal{I}_F^{-1}(\theta) = \text{As-cov}\{\hat{\theta}_{ML}\}$ corresponde a la inversa de la **matriz de información de Fisher** (FIM) [1–3, 98] (ver

Teorema 2.2). Si bien dicho resultado es conocido (ver e.g. [1, Capítulos 7 y 9]), es necesario adecuarlo a la distribución dada por (4.2).

La matriz de información de Fisher (FIM), cuando la distribución del ruido corresponde a un modelo de mezcla de distribuciones gaussianas, ha sido usualmente obtenida mediante integración numérica [132, 133]. No obstante, para el sistema dado por (4.1) y bajo los supuestos de que la distribución del ruido corresponde a una mezcla de 2 gaussianas, i.e. $M = 2$ en (4.2), que las medias de ambas componentes son cero $\mu_1 = \mu_2 = 0$ y que las varianzas se encuentran dadas por $\sigma_1^2 = \sigma^2$ y $\sigma_2^2 = \lambda\sigma_1^2$ con $\lambda > 1$, la Matriz de información de Fisher se encuentra en [132]. Sin embargo, en [134] se presenta una expresión cerrada para la FIM en términos de la función trascendente de Lerch [135], la función hipergeométrica gaussiana y una función hipergeométrica generalizada [136].

Lema 4.2. *La covarianza asintótica del error de estimación del estimador ML, $As-cov\{\hat{\theta}_{ML}\}$, para el sistema (4.1) se encuentra dada por:*

$$As-cov\{\hat{\theta}_{ML}\} = \frac{\sigma^2}{N\pi} \left[c + r - r(\lambda - 1)^2 + r\Phi\left(-\zeta\lambda, 2, \frac{1}{\lambda - 1}\right) \right]^{-1}, \quad (4.11)$$

donde $c = \frac{\alpha_1}{\pi}$, $r = \frac{\alpha_2}{\pi\lambda}$, $\zeta = \frac{1}{\alpha_2} - 1$ y $\Phi(z, q, a)$ es la función trascendental de Lerch dada por

$$\Phi(z, q, a) = \frac{1}{1 - z} \sum_{i=0}^{\infty} \left(\frac{-z}{1 - z} \right)^i \sum_{l=0}^{\infty} (-1)^l \binom{i}{l} (a + l)^{-q}, \quad (4.12)$$

lo cual se cumple para todo número complejo q y z con $\text{Re}[z] < 1/2$ [137].

Demostración 4.2. Ver [134, Teorema 2.1]. □

Observación 4.1. Usualmente el análisis asintótico para el estimador de ML ha sido abordado para sistemas lineales con ruido blanco Gaussiano (ver [1, 138]). No obstante, para el caso de una distribución del ruido dada por un Modelo de Mezcla gaussiana el análisis asintótico no es bien conocido, excepto para un caso particular de (4.1) con un GMM de dos componentes [134]. ▽

4.1.3. Comparación de precisión entre ambos estimadores

Para ilustrar la diferencia entre la covarianza del error de estimación del estimador de mínimos cuadrados LS y del estimador de máxima verosimilitud ML, considere el sistema dado por (4.1) con $\theta_0 = 1$, $\alpha_1 = 0,25$, $\alpha_2 = 0,75$, $\sigma_1^2 = 10$ y $\sigma_2^2 = 300$. La matriz de covarianza se obtiene para todo número de datos N entre 5 y 400. Se realizan 100 simulaciones de Monte Carlo (MC).

Los resultados de la simulación se pueden apreciar en la Fig. 4.1, donde las covarianzas teóricas (RT) de los estimadores LS y ML se obtienen mediante (4.5) y (4.11) respectivamente, mientras que las covarianzas muestrales (RE) se obtienen a partir de la simulación numérica. Note que el estimador de ML presenta una menor covarianza del error de estimación, pese a que ambos enfoques convergen hacia el mismo valor a medida que el número de datos N aumenta (debido a la propiedad de consistencia de ambos estimadores).

Por ende, la principal ventaja de utilizar una mezcla de distribuciones gaussianas en la estimación mediante ML es obtener una menor covarianza del error de estimación cuando el número de muestras es limitada. Esto ilustra la relevancia de conocer, o aproximar, la distribución del ruido. En [12] se presenta una extensión de este análisis para sistemas autoregresivos (AR).

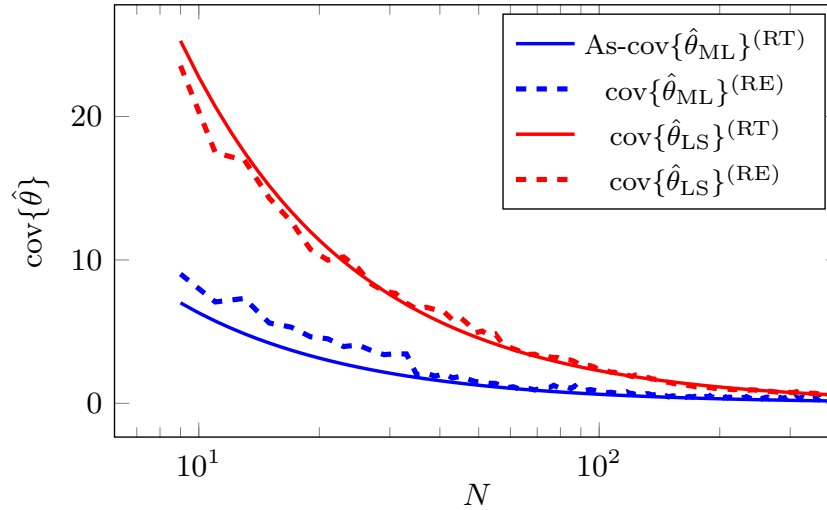


Fig. 4.1: Covarianzas del error de estimación para el ejemplo de estimar una constante cuando la distribución del ruido corresponde a un GMM de dos componentes.

4.2. Función de verosimilitud para ruido no-gaussiano

4.2.1. Modelo de mezcla de distribuciones gaussianas para aproximar ruido no-gaussiano

El modelo de mezcla de distribuciones gaussianas (GMM) ha sido previamente utilizado en la literatura para aproximar distribuciones no-gaussianas (ver e.g. [31, 139, 140]). A partir del **teorema de aproximación de Wiener**, es posible establecer que cualquier función de densidad de probabilidad (PDF) con soporte compacto se puede aproximar

mediante una suma finita de distribuciones Normales [14, 15]. La aproximación mediante combinación convexa de distribuciones gaussianas puede ser resumida de la siguiente manera:

Lema 4.3 (Aproximación de distribuciones mediante GMM).

Toda función de densidad de probabilidad de una variable aleatoria x de dimensión n , dada por $p(x|\beta)$, con soporte compacto puede ser aproximada tan cerca como se desee en el espacio $L_1(\mathbb{R}^n)$ por una distribución de la siguiente forma

$$p(x|\beta) \approx \sum_{j=1}^M \alpha_j \phi_j(x; \mu_j, \Sigma_j), \quad (4.13)$$

para algún $M \in \mathbb{N}$ y escalares positivos α_j tales que $\sum_{j=1}^M \alpha_j = 1$, los cuales se denominan pesos de la mezcla. $\phi_j(x; \mu_j, \Sigma_j)$ corresponde a una distribución gaussiana de dimensión n con vector de medias μ_j y matriz de covarianza Σ_j .

Demostración 4.3. Ver [14, Teorema 3] y [15, Página 224]. □

Observación 4.2. Debido a que mediante un modelo de mezcla de distribuciones gaussianas (GMM) es posible aproximar cualquier distribución con soporte compacto, este ha sido utilizado en diversos problemas tales como seguimiento [141], estimación Bayesiana [142], clasificación [34, 143], estimación de velocidad rotacional de estrellas [106] por mencionar algunos. ▽

Lecturas complementarias sobre la aproximación de distribuciones mediante combinaciones convexas de distribuciones gaussianas se pueden encontrar en [144] y [145].

4.2.2. Función de verosimilitud para mezcla de distribuciones gaussianas

Para el sistema dado por (4.1), se define el vector aumentado de parámetros a estimar como:

$$\beta = [\theta, \alpha_1, \mu_1, \sigma_1^2, \dots, \alpha_M, \mu_M, \sigma_M^2]. \quad (4.14)$$

Del mismo modo es posible definir β_0 como el vector de parámetros *verdaderos* del sistema. Al modelar el conjunto de mediciones, dado por $y_{1:N}$, como realizaciones de variables aleatorias, se tiene que

$$p(y_{1:N}|\beta) = \prod_{t=1}^N p(y_t|\beta). \quad (4.15)$$

Entonces, la **función de verosimilitud** $\mathcal{L}_N(\beta)$ puede ser escrita como

$$\mathcal{L}_N(\beta) = p(y_{1:N}|\beta) = \prod_{t=1}^N \sum_{j=1}^M \alpha_j \phi(y_t - \theta; \mu_j, \sigma_j^2). \quad (4.16)$$

Consecuentemente, el **logaritmo de la función de verosimilitud**, denotado por $\ell_N(\beta)$, se encuentra dado por

$$\ell_N(\beta) = \sum_{t=1}^N \log \left[\sum_{j=1}^M \alpha_j \phi(y_t - \theta; \mu_j, \sigma_j^2) \right]. \quad (4.17)$$

El problema de estimación mediante máxima verosimilitud (ML) es solucionado al maximizar el logaritmo de la función de verosimilitud dada por (4.17). La solución puede dificultarse a medida que aumenta el número de componentes de la mezcla M , dado que $\ell_N(\beta)$ puede contener *diversos máximos locales* y debido a la *gran dimensión del espacio de búsqueda*.

Observación 4.3. *Cada componente del GMM $\phi(\omega_t; \mu_j, \sigma_j^2)$ aumenta en 3 la dimensión del espacio de búsqueda, al agregar un peso α_j junto con la media μ_j y varianza σ_j^2 de la componente gaussiana adicional.* ∇

Considere el sistema dado por (4.1) con $\theta_0 = 1$, $N = 2000$ y un ruido ω_t distribuido como una mezcla de 4 gaussianas ($M = 4$) con pesos iguales $\alpha_j = 0,25$, varianzas iguales $\sigma_j^2 = 10$ y medias dadas por $\mu_1 = -5$, $\mu_2 = 5$, $\mu_3 = -15$ y $\mu_4 = 15$.

En la Fig. 4.2 es posible apreciar el logaritmo de la función de la verosimilitud para valores de 2 medias de la mezcla (μ_1 y μ_4) considerando que los parámetros restantes son conocidos.

En la Fig. 4.3 es posible observar la función logaritmo de la verosimilitud para θ y una de las medias (μ_4) manteniendo el resto de los parámetros en los valores verdaderos del sistema.

En ambas figuras se observan máximos locales, por ende es necesario utilizar un **algoritmo de optimización global** que permita sortear dichas dificultades. El algoritmo de búsqueda de patrones, más conocido como **Pattern Search** es utilizado para solucionar el problema de estimación dado por (4.17). Las razones por las cuales se utiliza dicho algoritmo, junto con una introducción a la optimización global, son tratadas en detalle en el **Anexo B**. La aparición de máximos locales también se presenta para los demás parámetros del modelo de mezcla de distribuciones gaussianas, es decir las covarianzas σ_j^2 y los pesos de la mezcla α_j .

Dificultades similares ocurren al extender el problema a otros escenarios como por ejemplo el de sistemas dinámicos, el cual es abordado posteriormente en la **Sección 4.3**.

Observación 4.4 (Unicidad de la solución).

La solución al problema de estimación mediante ML al utilizar un GMM para aproximar la distribución no-gaussiana del ruido **no es única**. Una permutación de las componentes de mezcla $\phi_j ; j \in [1, M]$ produce el mismo valor del logaritmo de la función de verosimilitud [45] (Ver Fig. 4.2). Debido a esto es necesario utilizar un método de optimización global. ∇

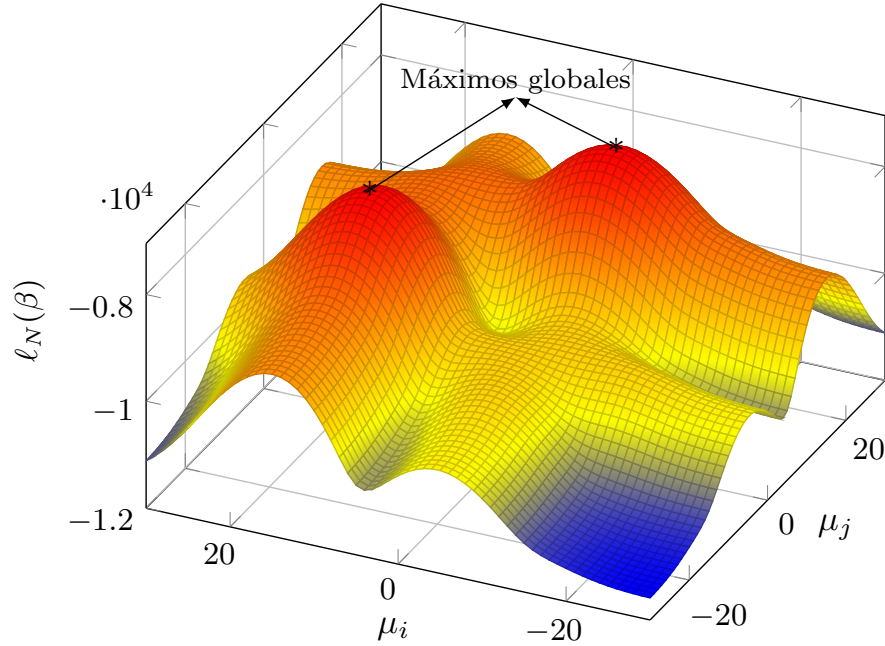


Fig. 4.2: Logaritmo de la función de verosimilitud para valores medios μ_i y μ_j con $i \neq j$ al utilizar GMM para aproximar la distribución del ruido.

4.3. Extensión a Sistemas Dinámicos Lineales

Considere un sistema lineal invariante en el tiempo genérico (revisar sistemas lineales invariantes en el tiempo en la página 15) dado por:

$$y_t = G_0(q^{-1})u_t + H_0(q^{-1})\omega_t, \quad (4.18)$$

donde q^{-1} corresponde al operador retardo de tiempo (i.e. $q^{-1}y_t = y_{t-1}$), $G_0(q^{-1})$ y $H_0(q^{-1})$ son funciones de transferencia estables, u_t corresponde a la entrada determinística del sistema y ω_t a ruido blanco con distribución no-gaussiana la cual puede ser aproximada mediante (4.13). De manera análoga al método del error de predicción PEM (ver

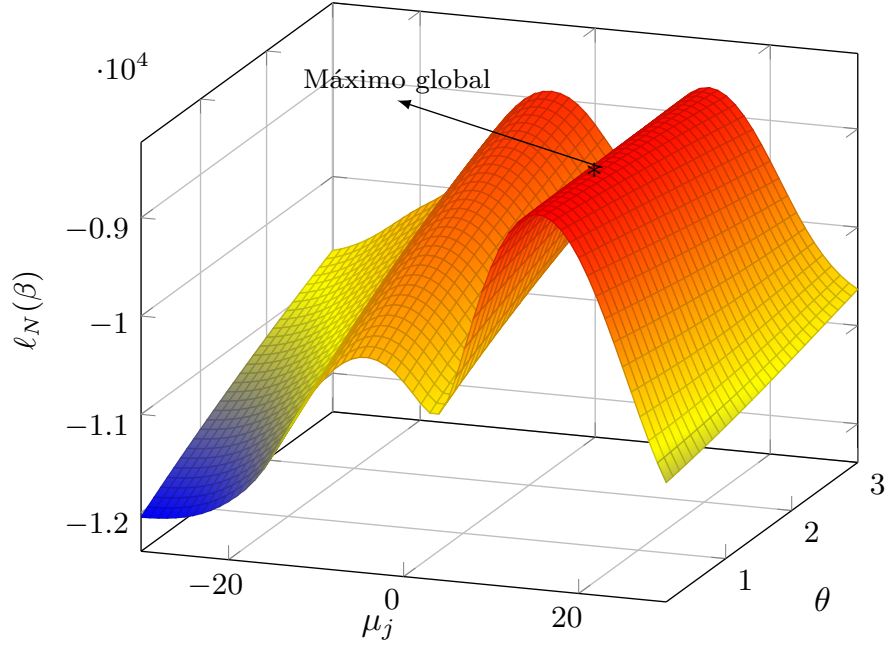


Fig. 4.3: Logaritmo de la función de verosimilitud para el valor medio μ_j del GMM y el parámetro θ .

Sección 2.4), se define el error de predicción, $\varepsilon_t(\theta)$, como [1, 2]:

$$\varepsilon_t(\theta) = \frac{1}{H(q^{-1}, \theta)} (y_t - G(q^{-1}, \theta) u_t). \quad (4.19)$$

Entonces, usando (4.17) para el sistema (4.18), y aproximando la distribución del ruido mediante un GMM (4.13), se obtiene:

$$\ell_N(\beta) = \sum_{t=1}^N \log \left[\sum_{j=1}^M \alpha_j \phi(\varepsilon_t(\theta); \mu_j, \sigma_j^2) \right]. \quad (4.20)$$

Luego, el **estimador de máxima verosimilitud** corresponde a:

$$\hat{\beta}_{\text{ML}} = \arg \max_{\beta} \ell_N(\beta) \quad \text{sujeto a} \quad \sum_{j=1}^M \alpha_j = 1 \quad (4.21)$$

Supuesto 4.1 (Submodelado).

Se asume que existe al menos un θ_0 tal que $H(q^{-1}, \theta_0) = H_0(q^{-1})$ y $G(q^{-1}, \theta_0) = G_0(q^{-1})$, i.e. el sistema no se encuentra *submodelado*.

Supuesto 4.2 (Condiciones de regularidad y fase mínima).

Se asume que el vector de parámetros β_0 , la entrada u_t y el ruido ω_t satisfacen condiciones

de regularidad de manera que la solución $\hat{\beta}_{\text{ML}}$ del problema de optimización dado por (4.21) converge (en probabilidad o de manera casi segura) hacia los parámetros del sistema β_0 . Adicionalmente, se asume que la función de transferencia $H(q^{-1}, \theta)$ es de fase mínima.

Observación 4.5. *El método de estimación mediante máxima verosimilitud dado por (4.21) posee el mismo problema que el método del error de predicción PEM ante ceros de fase no mínima de $H(q^{-1}, \theta)$, debido a que el predictor a un paso $\hat{y}_{t|t-1}$ se vuelve inestable. Este problema ha sido tema de investigación de varios autores (ver e.g. [42, 43]) y es abordado posteriormente en el **Capítulo 5**.* ∇

Observación 4.6. *La función logaritmo de la verosimilitud dada por (4.20) no considera las condiciones iniciales del sistema. No obstante las condiciones iniciales del sistema (4.18) pueden incluirse en θ mediante algunos cambios a la función de verosimilitud (ver e.g. [146]).* ∇

Observación 4.7. *Note que el funcional de costo en (4.20) corresponde al costo de PEM (dado por $N^{-1} \sum_{t=1}^N \varepsilon_t^2(\theta)$) cuando el número de componentes de la mezcla M es unitario, la media μ_1 del ruido es cero y $\alpha_1 = 1$ (ver **Ejemplo 2.2** y [1, 2]).* ∇

Observación 4.8 (Muestras de la distribución estimada).

En algunos casos resulta deseable obtener muestras a partir de la función de densidad de probabilidad estimada $\hat{p}(w_t)$ aproximada mediante un GMM (4.13). En dicho caso, las muestras se pueden generar a partir de M distribuciones normales $\{\phi_j(\mu_j, \Sigma_j)\}_{j=1}^M$ y la distribución categórica (también denominada distribución de Bernoulli generalizada) con $p(x = \phi_j) = \alpha_j$, donde α_j representa el peso de la mezcla asociado a la componente ϕ_j . Otra manera de obtener muestras es mediante el método de rebanadas (más conocido por su nombre en inglés, Slice Sampler) [147]. ∇

4.4. Ejemplos numéricos

En esta sección se presentan dos ejemplos numéricos con la finalidad de analizar el desempeño del estimador de máxima verosimilitud dado por (4.21), considerando un sistema dinámico LTI con fuente de ruido de distribución no-gaussiana. Considere que las observaciones (mediciones) $y_{1:N}$ son generadas a partir del siguiente *sistema real*:

$$y_t = \frac{b_1 q^{-1}}{1 + a_1 q^{-1} + a_2 q^{-1}} u_t + \frac{1 + c_1 q^{-1}}{1 + d_1 q^{-1}} \omega_t, \quad (4.22)$$

con $b_1 = 1$, $a_1 = -0,5$, $a_2 = 0,2$, $c_1 = 0,1$ y $d_1 = -0,95$. La señal de entrada se encuentra dada por $u_t \sim \mathcal{N}(0, \sigma_u^2)$ con $\sigma_u^2 = 10$, mientras que el ruido ω_t posee una distribución no-gaussiana.

El vector aumentado de parámetros a estimar esta dado por $\beta = \{\theta, \eta\}$, con $\theta = \{b_1, a_1, a_2, c_1, d_1\}$ y $\eta = \{\alpha_j, \mu_j, \sigma_j^2\}_{j=1}^M$. El algoritmo de búsqueda de patrones, más conocido como Pattern Search (ver **Sección B.5**) [148] es utilizado para solucionar el problema de optimización dado por (4.21). Las razones por las cuales se utiliza dicho algoritmo de optimización global se encuentran detalladas en la **Sección B.6**.

4.4.1. Configuración de las simulaciones

La configuración de la simulación utilizada en ambos ejemplos es la siguiente:

- (1) Los valores iniciales para el modelo θ se encuentran dados por la estimación obtenida mediante PEM (ver **Sección 2.4**), es decir $\hat{\theta}^{(PEM)}$.
- (2) Se consideran tres aproximaciones mediante el GMM, con $M = 3, 5, 7$ componentes.
- (3) La inicialización del GMM se encuentra dado por la covarianza muestral del error de predicción de PEM $\varepsilon_t^{(PEM)}$ para las covarianzas $\{\sigma_j^2\}_{j=1}^M$. Los pesos de la mezcla α_j se eligen iguales tal que $\sum_{j=1}^M \alpha_j = 1$. Las medias de los componentes de la mezcla, denotadas por $\{\mu_j\}_{j=1}^M$, se encuentran equiespaciadas entre el valor máximo y mínimo del error de predicción de PEM, $\varepsilon_t^{(PEM)}$.
- (4) Se consideraron cinco longitudes de datos, dadas por $N = \{100, 500, 1000, 5000, 10000\}$.
- (5) El número de simulaciones de Monte-Carlo es de 100 para cada método y longitud de datos.
- (6) El criterio de detención del algoritmo de búsqueda se encuentra dado por

$$\left| \ell_{m-1}(\hat{\beta}) - \ell_m(\hat{\beta}) \right| / \left(1 + \left| \ell_m(\hat{\beta}) \right| \right) \leq 10^{-6},$$

donde m corresponde al número de iteraciones, o se alcanza el máximo de evaluaciones del funcional de costo (5000) o se logra el número máximo de iteraciones (1000).

4.4.2. Criterios de desempeño

Con la finalidad de comparar el desempeño del estimador ML en (4.21) con PEM² [1], se consideran los siguientes criterios:

²La solución de PEM fue obtenida utilizando el algoritmo de optimización global sobre el funcional de costo (4.20) con $M = 1$, $\alpha_1 = 1$ y $\mu_1 = 0$.

1. La **covarianza del error de estimación**, la cual se encuentra dada por

$$P_{\hat{\theta}} \triangleq \mathcal{E} \left\{ (\hat{\theta} - \theta_0)(\hat{\theta} - \theta_0)^T \right\}. \quad (4.23)$$

2. La **norma ℓ_1 del error de aproximación de la PDF**, dado por $\epsilon(\eta)$, se define de la siguiente manera

$$\|\epsilon(\eta)\|_1 = \left\| p(\omega_t)^{(\text{Real})} - \hat{p}(\omega_t)^{(\text{GMM})} \right\|_1, \quad (4.24)$$

donde $p(\omega_t)^{(\text{Real})}$ corresponde a la distribución *real* del ruido, mientras que $\hat{p}(\omega_t)^{(\text{GMM})}$ corresponde al promedio de las distribuciones estimadas para todas las simulaciones de MC.

3. El **tiempo medio de ejecución** t_e [s] de la estimación de los parámetros del sistema dinámico y la aproximación de la distribución del ruido mediante un modelo de mezcla gaussiana GMM.

El objetivo principal del análisis es mostrar que el desempeño de la estimación mejora al aumentar el número de componentes gaussianas M del GMM.

Observación 4.9. *En este sentido, un criterio de selección del modelo podría utilizarse para elegir el número de componentes de la mezcla M (ver e.g. [149]), no obstante no es el objetivo de este trabajo.* ∇

4.4.3. Ejemplo 1: Sistema dinámico con ruido de distribución uniforme

Considere el sistema definido por (4.22) con ω_t uniformemente distribuido en el intervalo $[-2, 2]$, i.e. $\omega_t \sim \mathcal{U}[-2, 2]$. Adicionalmente considere la configuración de la simulación descrita en la página 56.

A continuación en las Figs. 4.4 a 4.8 y Tablas 4.1 a 4.5 es posible apreciar los resultados obtenidos a partir de la simulación de MC para el sistema dinámico con ruido distribuido de manera uniforme para diferentes números de datos N , cuando se estiman simultáneamente los parámetros del sistema y los parámetros que aproximan la distribución del ruido mediante el método de máxima verosimilitud ML. En dichas figuras $p(\omega_t)^{\text{GMM}}$ corresponde al **promedio de las distribuciones estimadas** en la simulación de MC al utilizar un GMM para aproximar la **distribución del ruido** dada por $p(\omega_t)^{\text{Real}}$. El área sombreada de color gris presente en las figuras representa la **región de incertidumbre** (RI), i.e. el área donde se encuentran todas las aproximaciones de la función de densidad de probabilidad PDF para la simulación de Monte Carlo.

En las tablas previamente mencionadas se presentan las estadísticas de la simulación de Monte-Carlo utilizando los criterios de desempeño definidos en la página 56. Es decir

$P_{\hat{\theta}}$, $\|\epsilon(\eta)\|_1$ y t_e [s], correspondientes a la varianza media del error de estimación, la norma ℓ_1 del error de aproximación de la densidad del ruido y el tiempo medio de ejecución. Además en dichas tablas se aprecian $\mu_{\hat{\theta}_j}$ y $P_{\hat{\theta}_j}$ que representan la estimación media del parámetro j en las simulaciones de MC y la varianza media del error de estimación para el parámetro j .

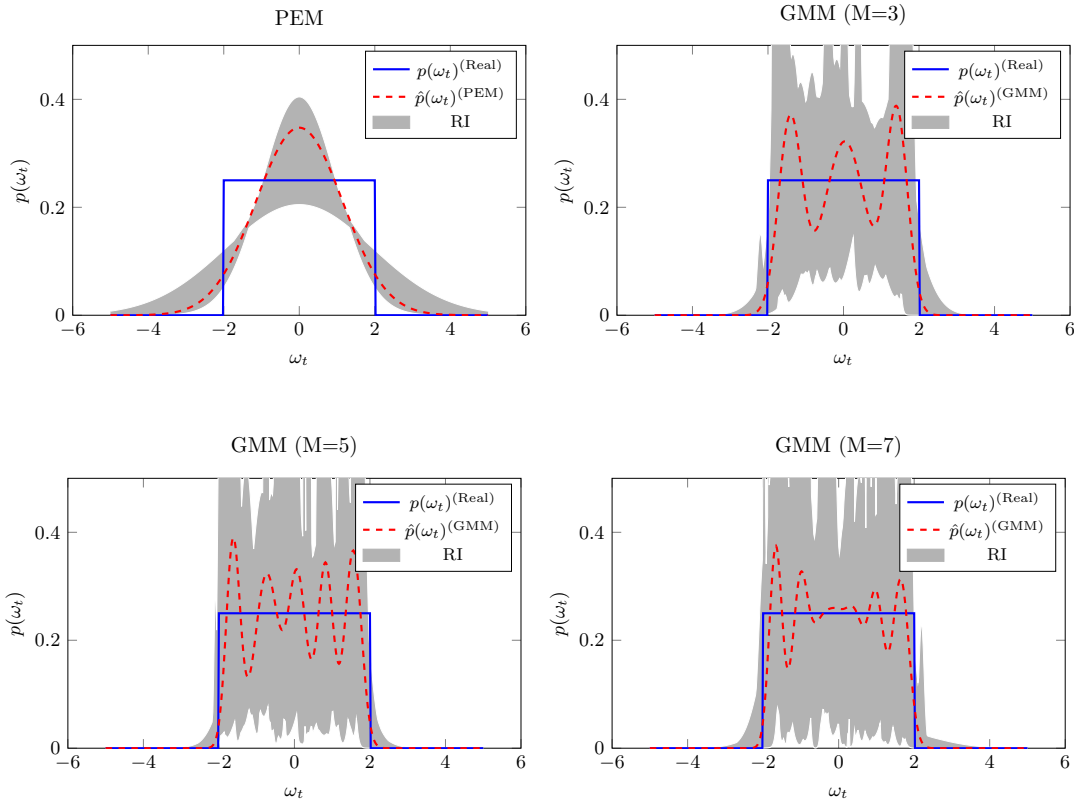
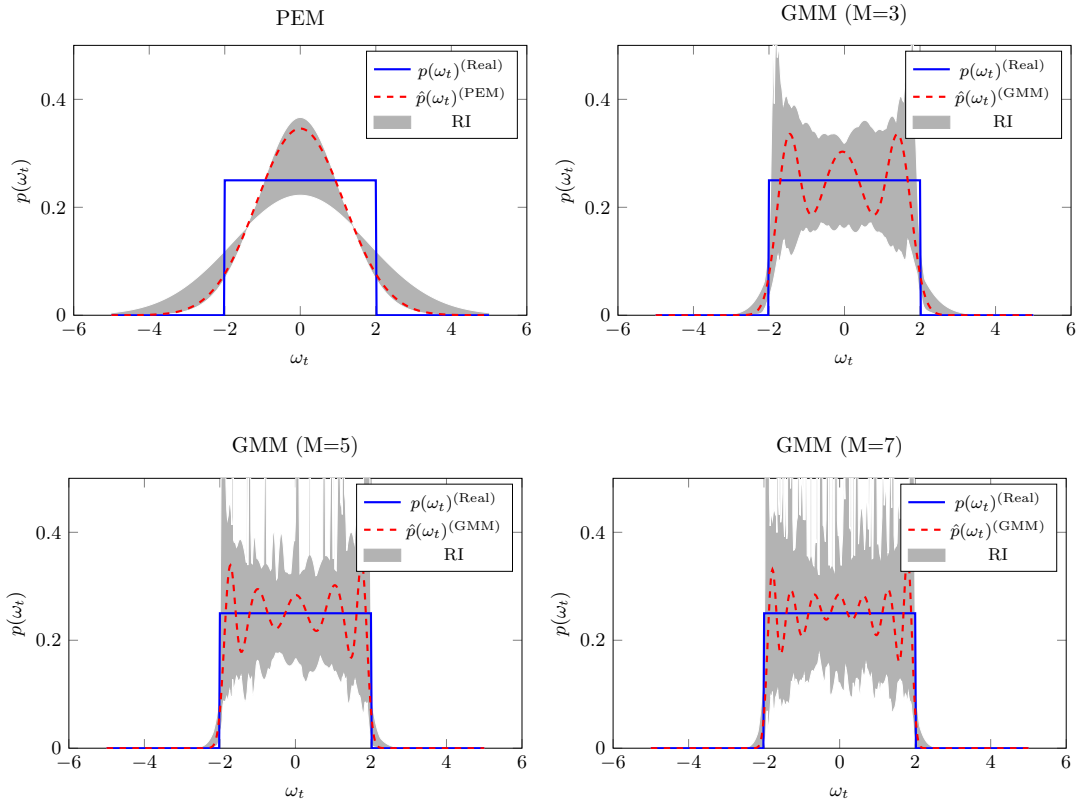
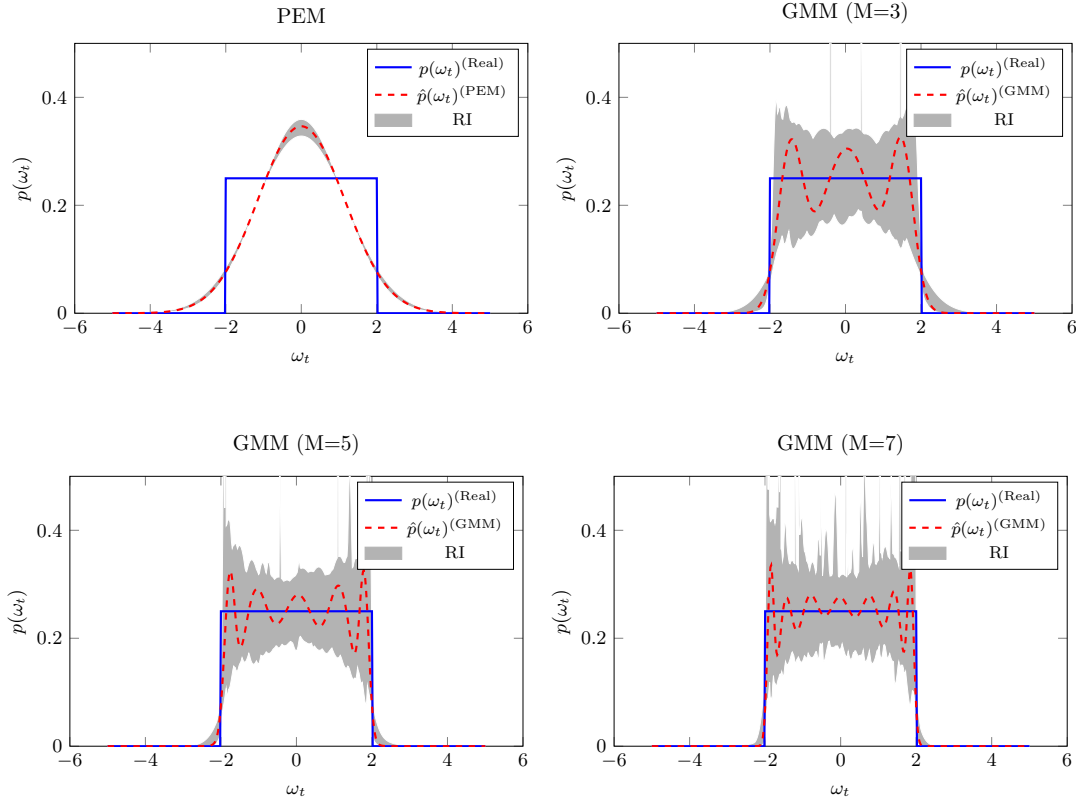


Fig. 4.4: Aproximaciones de la distribución uniforme del ruido al utilizar un GMM para el Ejemplo 1 con $N = 100$ datos.

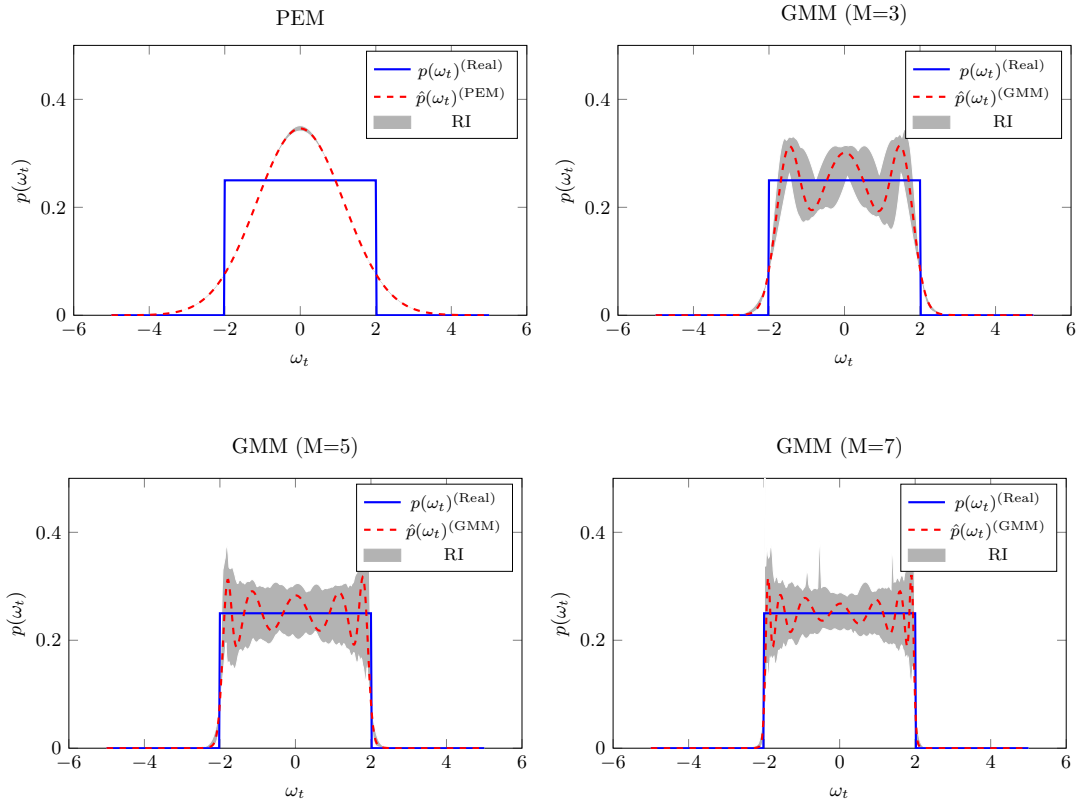
Método	$\mu_{\hat{\theta}_1}$ $P_{\hat{\theta}_1}$	$\mu_{\hat{\theta}_2}$ $P_{\hat{\theta}_2}$	$\mu_{\hat{\theta}_3}$ $P_{\hat{\theta}_3}$	$\mu_{\hat{\theta}_4}$ $P_{\hat{\theta}_4}$	$\mu_{\hat{\theta}_5}$ $P_{\hat{\theta}_5}$	$P_{\hat{\theta}}$	$\ \epsilon(\eta)\ _1$	t_e [s]
PEM	-0,5010	0,2016	0,9992	0,1398	-0,8910	0,081798	198.6788	1.7137
	0,01173	0,01159	0,01434	0,21634	0,17386			
GMM (M=3)	-0,5007	0,2009	1,0023	0,1120	-0,9324	0,009436	148.5049	9.7719
	0,00985	0,00881	0,01088	0,08192	0,04542			
GMM (M=5)	-0,4997	0,1999	1,0013	0,1106	-0,9442	0,006838	135.9757	21.6348
	0,01016	0,00871	0,01012	0,07168	0,03658			
GMM (M=7)	-0,4995	0,2010	0,9999	0,0895	-0,9393	0,007944	86.2663	33.9000
	0,00902	0,01037	0,00988	0,07533	0,04283			

Tabla 4.1: Estadísticas de la simulación de Monte-Carlo para el Ejemplo 1 con $N = 100$.Fig. 4.5: Aproximaciones de la distribución uniforme del ruido al utilizar un GMM para el Ejemplo 1 con $N = 500$ datos.

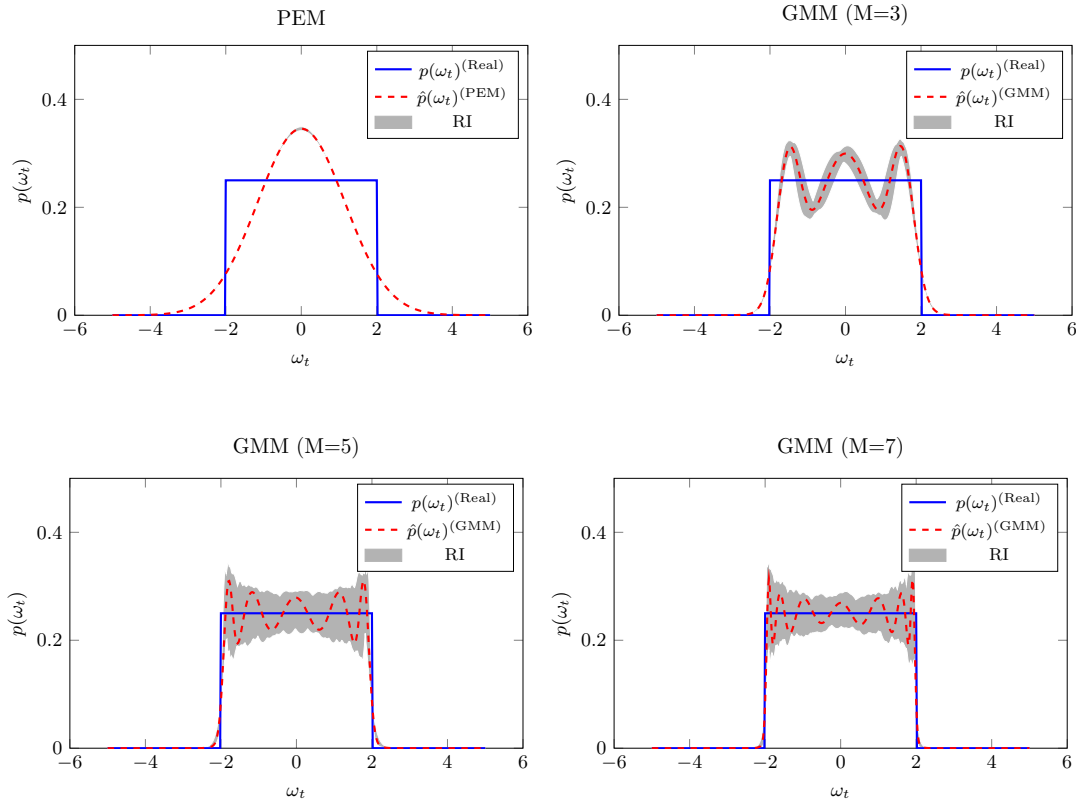
Método	$\mu_{\hat{\theta}_1}$ $P_{\hat{\theta}_1}$	$\mu_{\hat{\theta}_2}$ $P_{\hat{\theta}_2}$	$\mu_{\hat{\theta}_3}$ $P_{\hat{\theta}_3}$	$\mu_{\hat{\theta}_4}$ $P_{\hat{\theta}_4}$	$\mu_{\hat{\theta}_5}$ $P_{\hat{\theta}_5}$	$P_{\hat{\theta}}$	$\ \epsilon(\eta)\ _1$	$t_e[s]$
PEM	-0,4997	0,1997	1,0000	0,1180	-0,9365	0,013583	197.9417	1.4946
	0,00468	0,00457	0,00446	0,08603	0,07579			
GMM (M=3)	-0,4999	0,2002	1,0003	0,1031	-0,9482	0,000746	114.1453	7.2880
	0,00275	0,00265	0,00264	0,02511	0,00943			
GMM (M=5)	-0,5000	0,2003	0,9998	0,0988	-0,9502	0,000295	81.2055	19.0604
	0,00222	0,00223	0,00213	0,01490	0,00777			
GMM (M=7)	-0,4998	0,2004	1,0000	0,1004	-0,9496	0,000400	71.5985	35.7366
	0,00239	0,00248	0,00252	0,01767	0,00858			

Tabla 4.2: Estadísticas de la simulación de Monte-Carlo para el Ejemplo 1 con $N = 500$.Fig. 4.6: Aproximaciones de la distribución uniforme del ruido al utilizar un GMM para el Ejemplo 1 con $N = 1000$ datos.

Método	$\mu_{\hat{\theta}_1}$ $P_{\hat{\theta}_1}$	$\mu_{\hat{\theta}_2}$ $P_{\hat{\theta}_2}$	$\mu_{\hat{\theta}_3}$ $P_{\hat{\theta}_3}$	$\mu_{\hat{\theta}_4}$ $P_{\hat{\theta}_4}$	$\mu_{\hat{\theta}_5}$ $P_{\hat{\theta}_5}$	$P_{\hat{\theta}}$	$\ \epsilon(\eta)\ _1$	t_e [s]
PEM	-0,4993 0,00323	0,1998 0,00273	0,9994 0,00338	0,1002 0,03607	-0,9471 0,01139	0,001454	198.1862	1.1262
GMM (M=3)	-0,4998 0,00145	0,2000 0,00178	0,9998 0,00208	0,0984 0,01589	-0,9483 0,00600	0,000301	108.6302	7.4589
GMM (M=5)	-0,4999 0,00126	0,2001 0,00114	0,9997 0,00145	0,0993 0,01291	-0,9490 0,00483	0,000195	71.8924	17.8410
GMM (M=7)	-0,5002 0,00115	0,2001 0,00151	0,9997 0,00162	0,1000 0,00832	-0,9498 0,00401	0,000091	55.8285	33.5500

Tabla 4.3: Estadísticas de la simulación de Monte-Carlo para el Ejemplo 1 con $N = 1000$.Fig. 4.7: Aproximaciones de la distribución uniforme del ruido al utilizar un GMM para el Ejemplo 1 con $N = 5000$ datos.

Método	$\mu_{\hat{\theta}_1}$ $P_{\hat{\theta}_1}$	$\mu_{\hat{\theta}_2}$ $P_{\hat{\theta}_2}$	$\mu_{\hat{\theta}_3}$ $P_{\hat{\theta}_3}$	$\mu_{\hat{\theta}_4}$ $P_{\hat{\theta}_4}$	$\mu_{\hat{\theta}_5}$ $P_{\hat{\theta}_5}$	$P_{\hat{\theta}}$	$\ \epsilon(\eta)\ _1$	t_e [s]
PEM	-0,5001	0,2002	1,0000	0,0987	-0,9486	0,000241	197.9880	1.7012
	0,00147	0,00141	0,00153	0,01453	0,00465			
GMM M=(3)	-0,5001	0,2001	1,0000	0,0995	-0,9493	0,000068	102.2901	7.9552
	0,00072	0,00081	0,00088	0,00780	0,00227			
GMM M=(5)	-0,5001	0,2000	1,0000	0,0997	-0,9497	0,000026	67.1575	18.4397
	0,00050	0,00050	0,00057	0,00472	0,00157			
GMM M=(7)	-0,5000	0,2000	1,0000	0,0999	-0,9498	0,000013	46.5966	34.0591
	0,00036	0,00035	0,00036	0,00332	0,00123			

Tabla 4.4: Estadísticas de la simulación de Monte-Carlo para el Ejemplo 1 con $N = 5000$.Fig. 4.8: Aproximaciones de la distribución uniforme del ruido al utilizar un GMM para el Ejemplo 1 con $N = 10000$ datos.

Método	$\mu_{\hat{\theta}_1}$ $P_{\hat{\theta}_1}$	$\mu_{\hat{\theta}_2}$ $P_{\hat{\theta}_2}$	$\mu_{\hat{\theta}_3}$ $P_{\hat{\theta}_3}$	$\mu_{\hat{\theta}_4}$ $P_{\hat{\theta}_4}$	$\mu_{\hat{\theta}_5}$ $P_{\hat{\theta}_5}$	$P_{\hat{\theta}}$	$ \epsilon(\eta) _1$	t_e [s]
PEM	-0,5002	0,1998	1,0002	0,1004	-0,9497	0,000053	197.8207	1.7317
	0,00092	0,00094	0,00112	0,00637	0,00303			
GMM M=(3)	-0,5000	0,2000	1,0000	0,1006	-0,9499	0,000022	101.0719	8.6374
	0,00048	0,00057	0,00052	0,00431	0,00145			
GMM M=(5)	-0,5001	0,1999	1,0000	0,0997	-0,9501	0,000010	64.6547	19.9937
	0,00037	0,00042	0,00034	0,00282	0,00115			
GMM M=(7)	-0,5000	0,2000	1,0000	0,0998	-0,9501	0,000004	46.2651	36.1523
	0,00023	0,00028	0,00021	0,00195	0,00070			

Tabla 4.5: Estadísticas de la simulación de Monte-Carlo para el Ejemplo 1 con $N = 10000$.

4.4.4. Ejemplo 2: Sistema dinámico con ruido de distribución χ^2

En este ejemplo se considera que el ruido ω_t en (4.22) posee una distribución *Chi-cuadrado*, i.e. $\omega_t \sim \chi^2(\kappa)$ con $\kappa = 2$. El vector aumentado de parámetros a estimar se encuentra dado por $\beta = \{\theta, \eta\}$, con $\theta = \{b_1, a_1, a_2, c_1, d_1\}$ y $\eta = \{\alpha_j, \mu_j, \sigma_j^2\}_{j=1}^M$. Los parámetros reales del sistema se encuentran dados por $b_1 = 1$, $a_1 = -0,5$, $a_2 = 0,2$, $c_1 = 0,1$ y $d_1 = -0,95$. La configuración de la simulación corresponde a la previamente descrita en la **Subsección 4.4.1**. En las Figs. 4.9 a 4.13 y Tablas 4.6 a 4.10 se pueden revisar los resultados obtenidos de la simulación de MC.

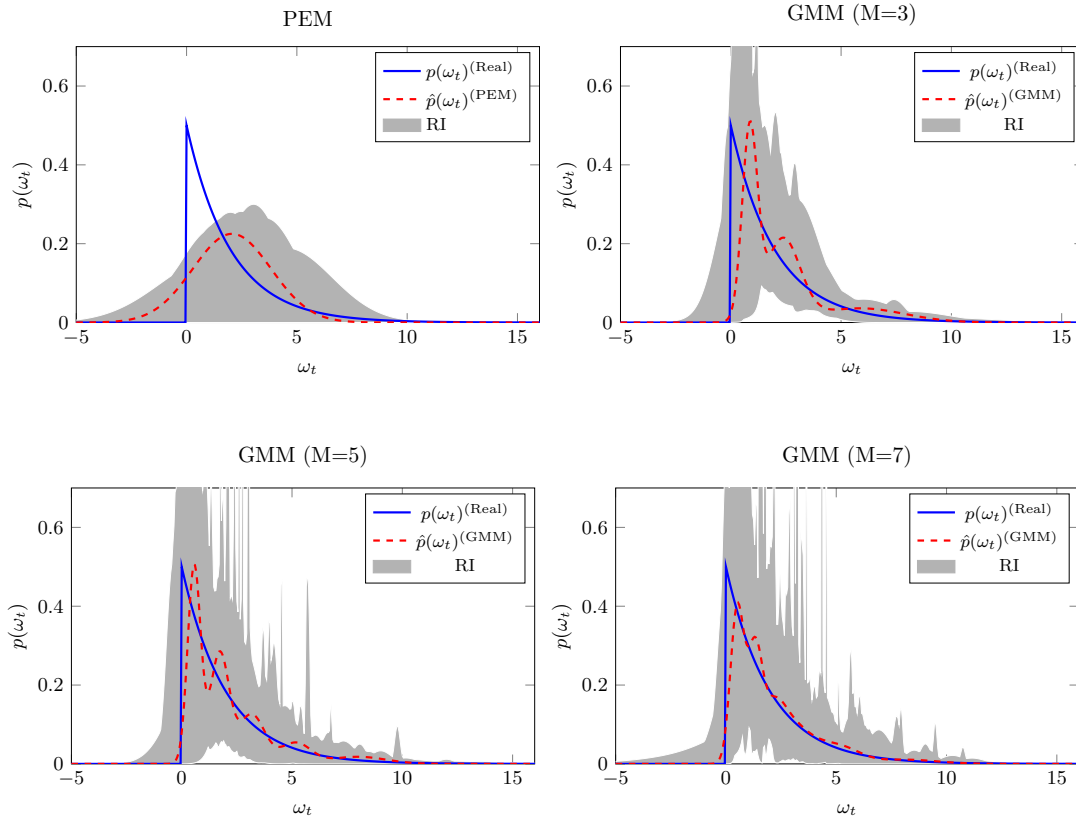


Fig. 4.9: Aproximaciones de la distribución *Chi-cuadrado* del ruido al utilizar un GMM para el Ejemplo 2 con $N = 100$ datos.

Método	$\mu_{\hat{\theta}_1}$ $P_{\hat{\theta}_1}$	$\mu_{\hat{\theta}_2}$ $P_{\hat{\theta}_2}$	$\mu_{\hat{\theta}_3}$ $P_{\hat{\theta}_3}$	$\mu_{\hat{\theta}_4}$ $P_{\hat{\theta}_4}$	$\mu_{\hat{\theta}_5}$ $P_{\hat{\theta}_5}$	$P_{\hat{\theta}}$	$\ \epsilon(\eta)\ _1$	t_e [s]
PEM	-0,4989 0,019878	0,1954 0,029737	0,9954 0,02803	0,0874 0,088656	-0,9421 0,024142	0,010668	125.5257	2.7313
GMM (M=3)	-0,4979 0,013807	0,1971 0,010717	0,9991 0,011886	0,1137 0,094975	-0,9422 0,015256	0,009866	83.2868	7.9366
GMM (M=5)	-0,4990 0,009869	0,1979 0,011965	1,0014 0,009303	0,1074 0,070433	-0,9479 0,010971	0,005421	57.8030	18.4965
GMM (M=7)	-0,4985 0,011516	0,1985 0,01576	1,0003 0,016607	0,1058 0,056948	-0,9464 0,010152	0,004014	38.2280	29.6279

Tabla 4.6: Estadísticas de la simulación de Monte-Carlo para el Ejemplo 2 con $N = 100$.

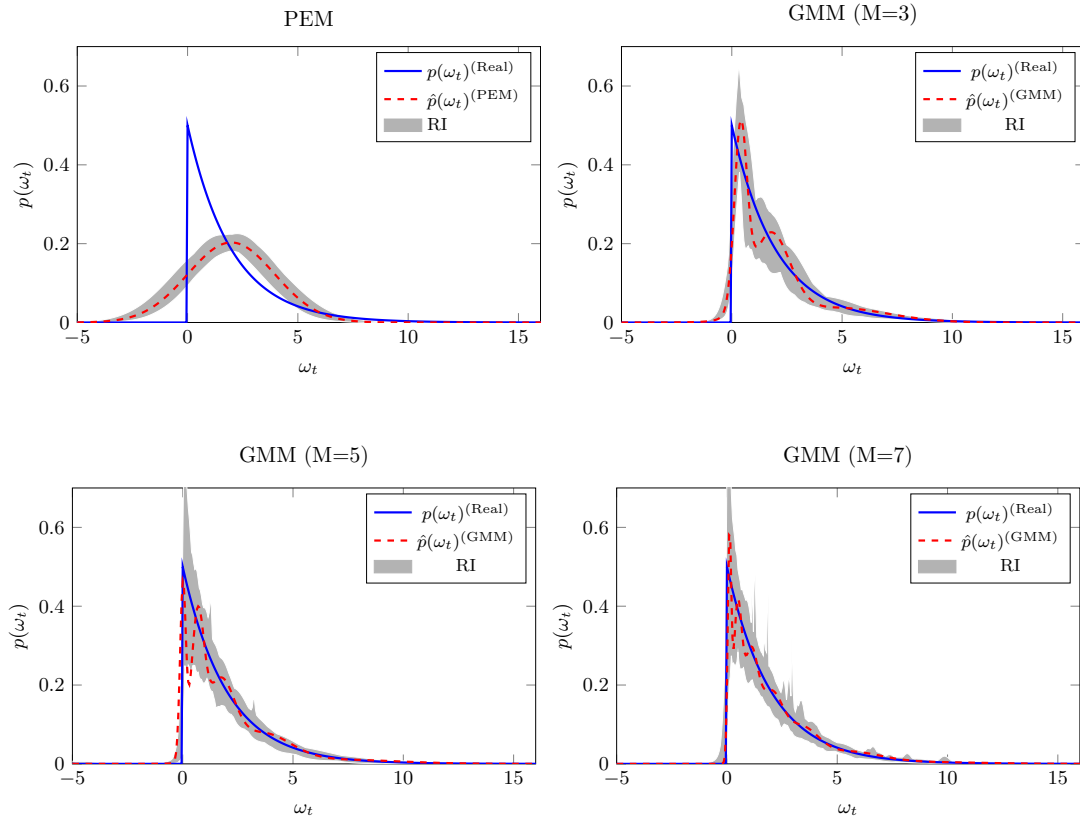


Fig. 4.10: Aproximaciones de la distribución *Chi-cuadrado* del ruido al utilizar un GMM para el Ejemplo 2 con $N = 500$ datos.

Método	$\mu_{\hat{\theta}_1}$ $P_{\hat{\theta}_1}$	$\mu_{\hat{\theta}_2}$ $P_{\hat{\theta}_2}$	$\mu_{\hat{\theta}_3}$ $P_{\hat{\theta}_3}$	$\mu_{\hat{\theta}_4}$ $P_{\hat{\theta}_4}$	$\mu_{\hat{\theta}_5}$ $P_{\hat{\theta}_5}$	$P_{\hat{\theta}}$	$\ \epsilon(\eta)\ _1$	t_e [s]
PEM	-0,5006 0,00776	0,2005 0,00751	1,0014 0,00849	0,0964 0,02927	-0,9520 0,00712	0,001104	125.5542	3.4223
GMM (M=3)	-0,5006 0,00321	0,2003 0,00334	1,0002 0,00342	0,0997 0,01700	-0,9500 0,00453	0,000340	54.4225	7.9882
GMM (M=5)	-0,5001 0,00283	0,1999 0,00315	0,9995 0,00561	0,1004 0,01486	-0,9500 0,00392	0,000283	32.6325	17.1827
GMM (M=7)	-0,5000 0,00273	0,2004 0,00257	1,0000 0,00317	0,1008 0,01185	-0,9500 0,00276	0,000171	28.3671	31.6666

Tabla 4.7: Estadísticas de la simulación de Monte-Carlo para el Ejemplo 2 con $N = 500$.

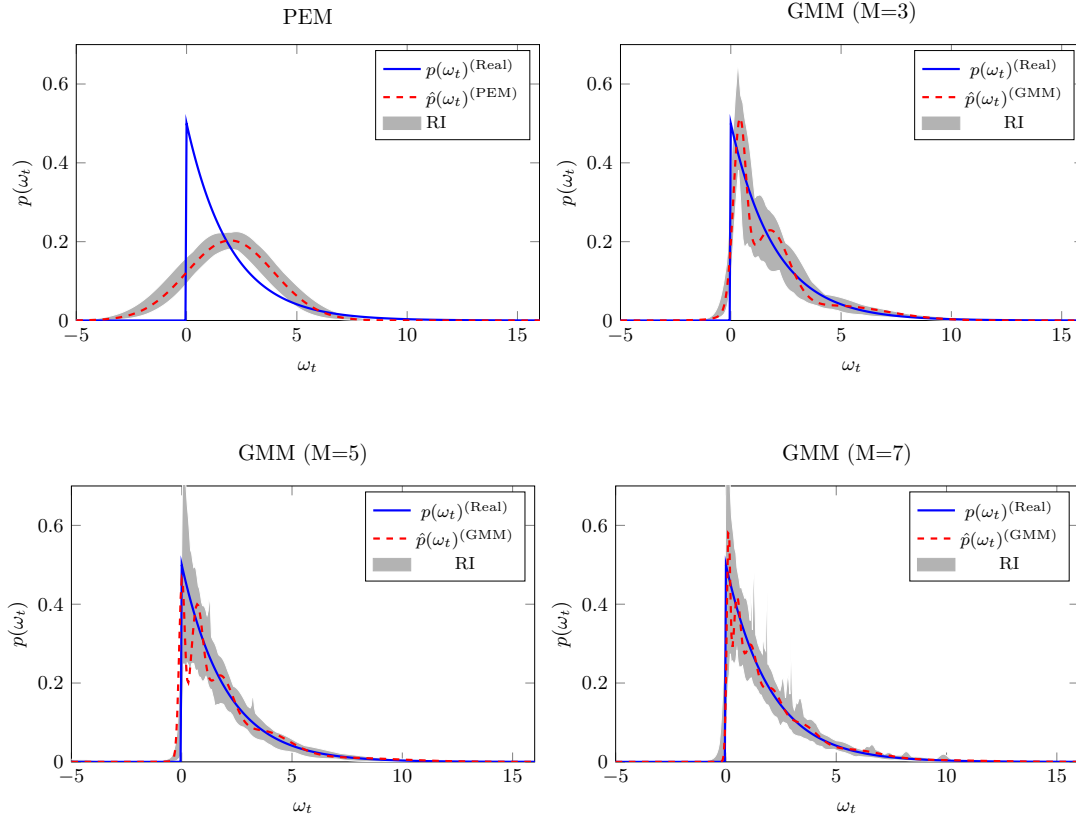


Fig. 4.11: Aproximaciones de la distribución *Chi-cuadrado* del ruido al utilizar un GMM para el Ejemplo 2 con $N = 1000$ datos.

Método	$\mu_{\hat{\theta}_1}$ $P_{\hat{\theta}_1}$	$\mu_{\hat{\theta}_2}$ $P_{\hat{\theta}_2}$	$\mu_{\hat{\theta}_3}$ $P_{\hat{\theta}_3}$	$\mu_{\hat{\theta}_4}$ $P_{\hat{\theta}_4}$	$\mu_{\hat{\theta}_5}$ $P_{\hat{\theta}_5}$	$P_{\hat{\theta}}$	$\ \epsilon(\eta)\ _1$	t_e [s]
PEM	-0,4996 0,00533	0,1994 0,00458	0,9998 0,00585	0,1005 0,01524	-0,9501 0,00672	0,000358	126.5728	3.5115
GMM (M=3)	-0,4999 0,00193	0,2000 0,00180	1,0000 0,00198	0,0990 0,00980	-0,9510 0,00171	0,000111	52.0947	9.0992
GMM (M=5)	-0,4999 0,00123	0,2000 0,00110	1,0000 0,00161	0,1004 0,00736	-0,9506 0,00115	0,000061	46.8232	16.8765
GMM (M=7)	-0,4999 0,00162	0,2000 0,00129	1,0001 0,00170	0,1006 0,00856	-0,9504 0,00134	0,000082	23.1480	31.4654

Tabla 4.8: Estadísticas de la simulación de Monte-Carlo para el Ejemplo 2 con $N = 1000$.

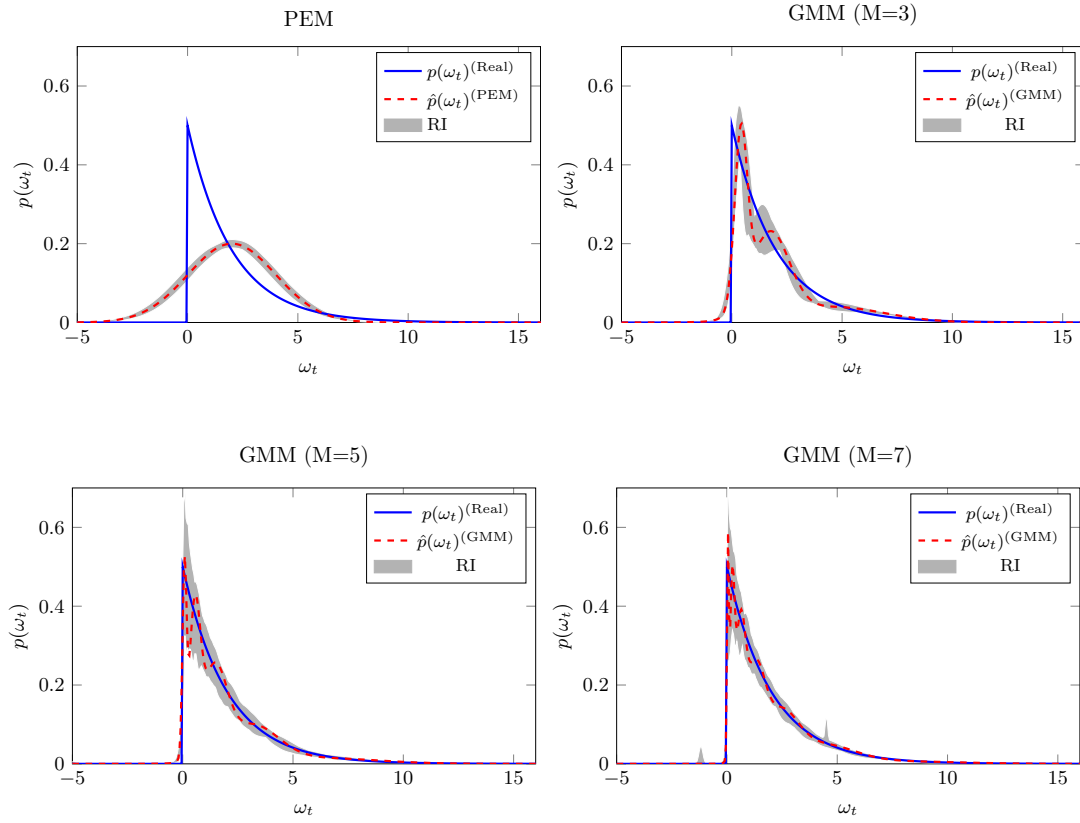


Fig. 4.12: Aproximaciones de la distribución *Chi-cuadrado* del ruido al utilizar un GMM para el Ejemplo 2 con $N = 5000$ datos.

Método	$\mu_{\hat{\theta}_1}$ $P_{\hat{\theta}_1}$	$\mu_{\hat{\theta}_2}$ $P_{\hat{\theta}_2}$	$\mu_{\hat{\theta}_3}$ $P_{\hat{\theta}_3}$	$\mu_{\hat{\theta}_4}$ $P_{\hat{\theta}_4}$	$\mu_{\hat{\theta}_5}$ $P_{\hat{\theta}_5}$	$P_{\hat{\theta}}$	$\ \epsilon(\eta)\ _1$	t_e [s]
PEM	-0,5001 0,00235	0,2002 0,00224	1,0001 0,00232	0,1013 0,01198	-0,9498 0,00308	0,000169	127.8471	3.8465
GMM (M=3)	-0,5000 0,00081	0,2000 0,00082	1,0000 0,00076	0,0995 0,00396	-0,9505 0,00102	0,000019	51.0048	8.6313
GMM (M=5)	-0,5000 0,00049	0,2000 0,00048	1,0000 0,00049	0,1002 0,00250	-0,9501 0,00061	0,000007	31.0971	17.4612
GMM (M=7)	-0,5000 0,00030	0,2000 0,00030	1,0000 0,00034	0,0999 0,00171	-0,9501 0,00036	0,000003	18.5576	31.4404

Tabla 4.9: Estadísticas de la simulación de Monte-Carlo para el Ejemplo 2 con $N = 5000$.

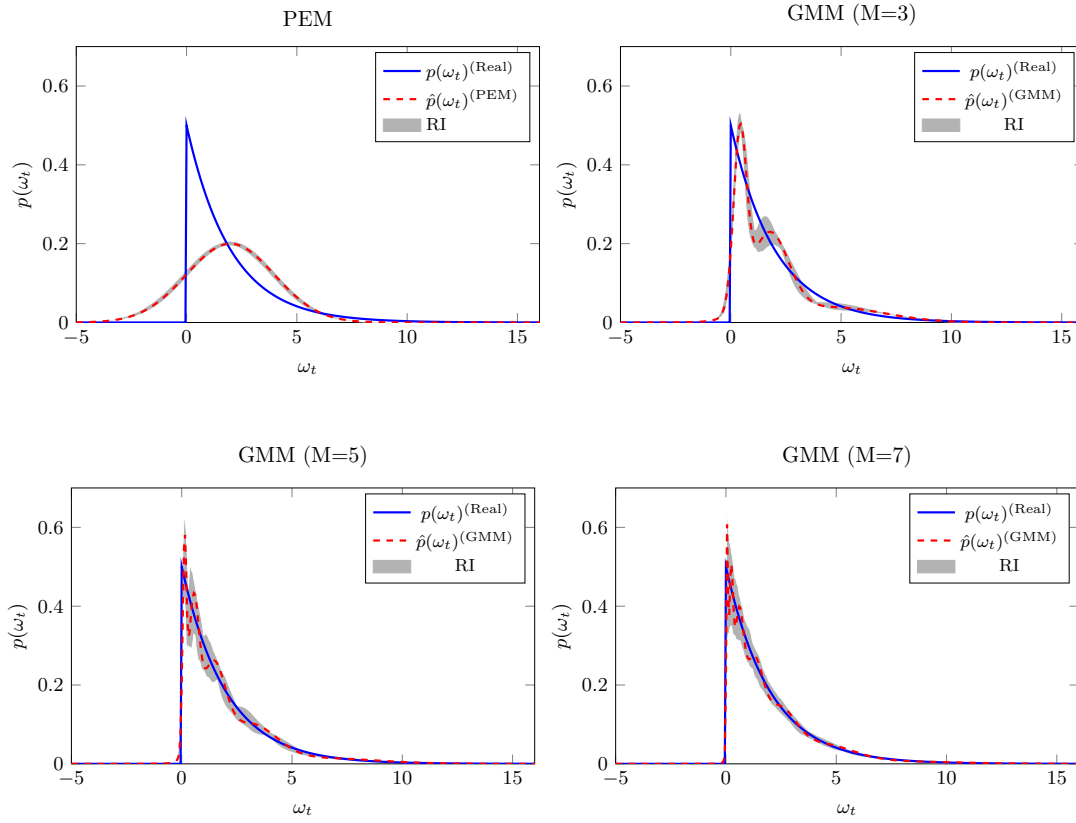


Fig. 4.13: Aproximaciones de la distribución *Chi-cuadrado* del ruido al utilizar un GMM para el Ejemplo 2 con $N = 10000$ datos.

Método	$\mu_{\hat{\theta}_1}$ $P_{\hat{\theta}_1}$	$\mu_{\hat{\theta}_2}$ $P_{\hat{\theta}_2}$	$\mu_{\hat{\theta}_3}$ $P_{\hat{\theta}_3}$	$\mu_{\hat{\theta}_4}$ $P_{\hat{\theta}_4}$	$\mu_{\hat{\theta}_5}$ $P_{\hat{\theta}_5}$	$P_{\hat{\theta}}$	$\ \epsilon(\eta)\ _1$	t_e [s]
PEM	-0,5002 0,00172	0,2000 0,00180	1,0000 0,00173	0,1014 0,00953	-0,9495 0,00286	0,000109	127.3036	4.1753
GMM (M=3)	-0,5000 0,00058	0,2000 0,00061	0,9999 0,00057	0,1003 0,00341	-0,9505 0,00081	0,000013	51.0223	10.1109
GMM (M=5)	-0,5000 0,00030	0,2000 0,00030	0,9999 0,00035	0,0996 0,00184	-0,9502 0,00041	0,000004	28.8270	19.6741
GMM (M=7)	-0,5000 0,00020	0,2000 0,00023	1,0000 0,00023	0,0999 0,00110	-0,9502 0,00026	0,000001	17.8605	33.0091

Tabla 4.10: Estadísticas de la simulación de Monte-Carlo para el Ejemplo 2 con $N = 10000$.

4.4.5. Resumen y análisis de los resultados

A continuación se resumen los criterios de desempeño obtenidos a partir de las simulaciones de Monte-Carlo para ambos ejemplos. Posteriormente se analizan los resultados expuestos.

N / Método	PEM		GMM (M=3)		GMM (M=5)		GMM (M=7)	
	$P_{\hat{\theta}}$	$\ \epsilon(\eta)\ _1$	$P_{\hat{\theta}}$	$\ \epsilon(\eta)\ _1$	$P_{\hat{\theta}}$	$\ \epsilon(\eta)\ _1$	$P_{\hat{\theta}}$	$\ \epsilon(\eta)\ _1$
100	$8,18 \times 10^{-2}$	$1,99 \times 10^2$	$9,44 \times 10^{-3}$	$1,49 \times 10^2$	$6,84 \times 10^{-3}$	$1,36 \times 10^2$	$7,94 \times 10^{-3}$	$8,63 \times 10^1$
500	$1,36 \times 10^{-2}$	$1,98 \times 10^2$	$7,46 \times 10^{-4}$	$1,14 \times 10^2$	$2,95 \times 10^{-4}$	$8,12 \times 10^1$	$4,00 \times 10^{-4}$	$7,16 \times 10^1$
1000	$1,45 \times 10^{-3}$	$1,98 \times 10^2$	$3,01 \times 10^{-4}$	$1,09 \times 10^2$	$1,95 \times 10^{-4}$	$7,19 \times 10^1$	$9,08 \times 10^{-5}$	$5,58 \times 10^1$
5000	$2,41 \times 10^{-4}$	$1,98 \times 10^2$	$6,80 \times 10^{-5}$	$1,02 \times 10^2$	$2,55 \times 10^{-5}$	$6,72 \times 10^1$	$1,29 \times 10^{-5}$	$4,66 \times 10^1$
10000	$5,26 \times 10^{-5}$	$1,98 \times 10^2$	$2,17 \times 10^{-5}$	$1,01 \times 10^2$	$9,70 \times 10^{-6}$	$6,47 \times 10^1$	$4,48 \times 10^{-6}$	$4,63 \times 10^1$

Tabla 4.11: Resumen de las estadísticas de la simulación de MC para el Ejemplo 1 (ver página 57).

N / Método	PEM		GMM (M=3)		GMM (M=5)		GMM (M=7)	
	$P_{\hat{\theta}}$	$\ \epsilon(\eta)\ _1$	$P_{\hat{\theta}}$	$\ \epsilon(\eta)\ _1$	$P_{\hat{\theta}}$	$\ \epsilon(\eta)\ _1$	$P_{\hat{\theta}}$	$\ \epsilon(\eta)\ _1$
100	$1,07 \times 10^{-2}$	$1,26 \times 10^2$	$9,87 \times 10^{-3}$	$8,33 \times 10^1$	$5,42 \times 10^{-3}$	$5,78 \times 10^1$	$4,01 \times 10^{-3}$	$3,82 \times 10^1$
500	$1,10 \times 10^{-3}$	$1,26 \times 10^2$	$3,40 \times 10^{-4}$	$5,44 \times 10^1$	$2,83 \times 10^{-4}$	$3,26 \times 10^1$	$1,71 \times 10^{-4}$	$2,84 \times 10^1$
1000	$3,58 \times 10^{-4}$	$1,27 \times 10^2$	$1,11 \times 10^{-4}$	$5,21 \times 10^1$	$6,08 \times 10^{-5}$	$4,68 \times 10^1$	$5,18 \times 10^{-5}$	$2,31 \times 10^1$
5000	$1,69 \times 10^{-4}$	$1,27 \times 10^2$	$1,89 \times 10^{-5}$	$5,10 \times 10^1$	$7,27 \times 10^{-6}$	$3,11 \times 10^1$	$3,33 \times 10^{-6}$	$1,86 \times 10^1$
10000	$1,10 \times 10^{-4}$	$1,27 \times 10^2$	$1,34 \times 10^{-5}$	$5,10 \times 10^1$	$4,04 \times 10^{-6}$	$2,88 \times 10^1$	$1,42 \times 10^{-6}$	$1,79 \times 10^1$

Tabla 4.12: Resumen de las estadísticas de la simulación de MC para el Ejemplo 2 (ver página 63).

N / Método	PEM		GMM (M=3)		GMM (M=5)		GMM (M=7)	
	$\mathcal{U}[-2, 2] [s]$	$\chi^2(2) [s]$	$\mathcal{U}[-2, 2] [s]$	$\chi^2(2) [s]$	$\mathcal{U}[-2, 2] [s]$	$\chi^2(2) [s]$	$\mathcal{U}[-2, 2] [s]$	$\chi^2(2) [s]$
100	1,7137	2,7313	9,7719	7,9366	21,6348	18,4965	33,9000	29,6279
500	1,4946	3,4223	7,2880	7,9882	19,0604	17,1827	35,7366	31,6666
1000	1,1262	3,5115	7,4589	8,0992	17,8410	16,8765	33,5500	31,4654
5000	1,7012	3,8465	7,9552	8,6313	18,4397	17,4612	34,0591	31,4404
10000	1,7317	4,1753	8,6374	10,1109	19,9937	19,6741	36,1523	33,0091

Tabla 4.13: Tiempos medios de ejecución $t_e [s]$ para los Ejemplos 1 y 2.

En las Tablas 4.11 y 4.12 se resumen los resultados, obtenidos a partir de la simulación de MC, del desempeño del algoritmo de estimación para diferentes longitudes de datos N .

La covarianza del error de estimación (4.23) es utilizada como criterio de desempeño de la estimación de los parámetros del sistema θ . Note que la covarianza del error de estimación $P_{\hat{\theta}}$ disminuye al incrementar el número de componentes de mezcla M en

ambos ejemplos, para una misma longitud de datos N .

La covarianza del error de estimación también se reduce al aumentar el número de datos N debido a la consistencia de los métodos utilizados (ver página 25), no obstante es posible apreciar que al incrementar el número de componentes M del GMM se obtiene una convergencia más rápida.

La norma ℓ_1 del error de aproximación de densidad de probabilidad $\epsilon(\eta)$ es usada para evaluar el desempeño de la aproximación de la PDF mediante un modelo de mezcla de distribuciones gaussianas GMM. Note que en ambos ejemplos y para un mismo número de datos N la aproximación de la densidad mejora a medida que se incrementa el número de componentes de mezcla M (dado que $\|\epsilon(\eta)\|_1$ disminuye). La aproximación también mejora al aumentar el número de datos disponibles N (al contar con más datos se reduce la incertidumbre). Resulta importante destacar que el error de aproximación de la función de densidad de probabilidad no mejora para el método del error de predicción PEM, debido a que solo puede ajustar los parámetros de una distribución normal.

El comportamiento previamente señalado se puede apreciar en las Figs. 4.4 a 4.8 para el primer ejemplo y en las Figs. 4.9 a 4.13 para el segundo ejemplo. En dichas figuras se puede apreciar que el promedio de las distribuciones estimadas, dada por $p(\omega_t)^{\text{GMM}}$, aproxima de mejor manera a la distribución del ruido $p(\omega_t)^{\text{Real}}$ a medida que aumenta el número de componentes de la mezcla M y/o el número de datos N . Además se observa que la región de incertidumbre RI presenta una gran dispersión respecto de la aproximación media de la distribución cuando se cuenta con un número pequeño de muestras. No obstante la región de incertidumbre se contrae a medida que aumenta el número de datos disponibles N . Esto es bastante intuitivo, puesto que al poseer mayor información disponible se obtienen resultados mejores y más confiables.

La Tabla 4.13 muestra el tiempo medio de ejecución t_e [s] de las simulaciones de Monte-Carlo en la estimación del sistema dinámico y aproximación de la distribución del ruido utilizando un GMM para los Ejemplos 1 y 2. En dicha tabla se puede apreciar el aumento gradual de la carga computacional al incrementar el número de componentes del GMM (y por ende la dimensión del espacio de búsqueda) o la cantidad de datos N para ambos ejemplos.

A partir de los resultados obtenidos es posible afirmar que el método de máxima verosimilitud propuesto (4.21), para estimar los parámetros del sistema y los parámetros que aproximan la distribución del ruido ω_t como un GMM de manera simultanea, mejora significativamente la estimación de los parámetros del sistema respecto del método

del error de predicción PEM en términos de la covarianza del error de estimación (en general al utilizar $M = 7$ componentes se obtiene al menos un orden de magnitud). Este resultado es más evidente para un número reducido de datos N disponibles, debido a las propiedades asintóticas de ambos métodos. Adicionalmente se obtiene una buena aproximación para la función de densidad de probabilidad del ruido ω_t , la cual mejora al aumentar el número de componentes de la mezcla M y/o el número de datos disponibles N .

IDENTIFICACIÓN MEDIANTE ML DE SISTEMAS DE FASE NO-MÍNIMA CON RUIDO GMM

En el **Capítulo 4** se presenta una estrategia para resolver el problema de identificación de un sistema lineal genérico (ver (4.18)) cuando la distribución del ruido es no-gaussiana, mediante el método de máxima verosimilitud. No obstante, el problema se encuentra restringido a sistemas de **fase mínima** para la función de transferencia asociada al ruido, es decir $H(q^{-1}, \theta)$. El problema de estimación, dado por (4.17), es resuelto mediante un algoritmo de optimización global (ver **Anexo B**). Dicho algoritmo presenta algunas desventajas, entre las cuales destacan la elevada carga computacional, lo cual implica un gran tiempo de ejecución y el ajuste de los parámetros del algoritmo de optimización que impactan considerablemente sobre su desempeño.

A continuación se presenta una extensión de los resultados obtenidos en el **Capítulo 4** al considerar sistemas con ceros de fase no-mínima (NMPZ), junto con utilizar el algoritmo de **maximización de la esperanza** (EM) (Ver **Sección A.2**). Para ello se describe en primer lugar el problema a solucionar y los supuestos asociados a este. Luego, se extiende la expresión del **error de predicción** con la finalidad de incluir sistemas de fase no-mínima (ver **Ejemplo 2.3**), resultado necesario para formular el problema de estimación mediante máxima verosimilitud.

Posteriormente se desarrolla un algoritmo basado en la maximización de la esperanza (EM) para resolver el problema de identificación. Finalmente se analiza el desempeño del método propuesto mediante ejemplos numéricos. Los resultados se contrastan con el método del error de predicción cuando el sistema es de fase mínima y con el método de los momentos para fase no-mínima.

5.1. Descripción del problema

Considere un sistema lineal genérico (ver Tabla 2.1) dado por

$$y_t = G(z^{-1}, \theta)u_t + H(z^{-1}, \theta)\omega_t, \quad (5.1)$$

donde z corresponde al operador retardo de tiempo ($z^{-1}u_t = u_{t-1}$) o la variable asociada a la transformada- Z , mientras que u_t corresponde a la señal determinística de entrada del sistema, y_t a la salida y ω_t corresponde a ruido blanco de distribución no-gaussiana con media igual a cero. A partir de los resultados previamente obtenidos en el **Capítulo 4** supondremos que la distribución de ω_t se encuentra dada por el siguiente modelo de mezcla de distribuciones gaussianas

$$p(\omega_t) = \sum_{i=1}^M \alpha_i \phi(\omega_t; \mu_i, \Sigma_i), \quad (5.2)$$

con $M \geq 2$ ($M \in \mathbb{N}$), $\alpha_i > 0$, $\sum_{i=1}^M \alpha_i = 1$ y $\phi(\omega_t; \mu_i, \Sigma_i)$ representa una distribución gaussiana con media μ_i y covarianza $\Sigma_i > 0$.

El **Lema 4.3** indica que en caso de que ω_t posea una distribución diferente, esta puede ser aproximada arbitrariamente cerca por el modelo de mezcla de distribuciones gaussianas (5.2) para algún $M \geq 2$ ($M \in \mathbb{N}$).

Considere que la función de transferencia $G(z^{-1}, \theta)$ del sistema dado por (5.1) es estable, mientras que $H(z^{-1}, \theta)$ corresponde a una función de transferencia que contiene ceros de fase mínima y/o no-mínima. Ambas funciones de transferencia se encuentran parametrizadas por θ y se encuentran dadas por

$$G(z^{-1}, \theta) = \frac{B(z^{-1}, \theta)}{A(z^{-1}, \theta)} H(z^{-1}, \theta) = \frac{C(z^{-1}, \theta)}{D(z^{-1}, \theta)}, \quad (5.3)$$

con

$$A(z^{-1}, \theta) = 1 + a_1 z^{-1} + \dots + a_{n_a} z^{-n_a}, \quad (5.4a)$$

$$B(z^{-1}, \theta) = b_1 z^{-1} + \dots + b_{n_b} z^{-n_b}, \quad (5.4b)$$

$$C(z^{-1}, \theta) = 1 + c_1 z^{-1} + \dots + c_{n_c} z^{-n_c}, \quad (5.4c)$$

$$D(z^{-1}, \theta) = 1 + d_1 z^{-1} + \dots + d_{n_d} z^{-n_d} \quad (5.4d)$$

Dado que el polinomio $C(z^{-1}, \theta)$ puede contener ceros de fase no-mínima (NMPZ), considere que se encuentra dado por $C(z^{-1}, \theta) = C_s(z^{-1}, \theta)C_u(z^{-1}, \theta)$, donde

$$C_s(z^{-1}, \theta) = 1 + c_{s_1} z^{-1} + \dots + c_{s_r} z^{-r}, \quad (5.5)$$

$$C_u(z^{-1}, \theta) = 1 + c_{u_1} z^{-1} + \dots + c_{u_p} z^{-p}, \quad (5.6)$$

con $r + p = n_c$. $C_s(z^{-1}, \theta)$ representa la parte de fase mínima de $C(z^{-1}, \theta)$, i.e. todas las raíces de $C_s(z^{-1}, \theta)$ se encuentran dentro del círculo unitario, mientras que $C_u(z^{-1}, \theta)$ contiene la parte de fase no-mínima, es decir que todas las raíces de $C_u(z^{-1}, \theta)$ se encuentran fuera del círculo unitario. Resulta importante destacar que los coeficientes de los polinomios descritos en (5.4) son reales. Por ende, en caso de existir raíces complejas en $C_u(z^{-1}, \theta)$ o $C_s(z^{-1}, \theta)$ deben encontrarse en pares conjugados.

5.1.1. Supuestos

El problema en el cual estamos interesados es estimar el vector aumentado de parámetros dado por $\beta = [\theta, \gamma]$, donde γ corresponde al vector de parámetros asociados al modelo de mezcla de distribuciones gaussianas GMM del ruido ω_t , mientras que θ contiene los parámetros del sistema, es decir los coeficientes de los polinomios descritos en (5.4). Del mismo modo es posible definir β_0 como el vector de parámetros verdaderos del sistema.

A continuación se presentan los supuestos asociados al problema previamente descrito.

Supuesto 5.1 (Lazo abierto). El sistema dado por (5.1) se encuentra en lazo abierto y u_t corresponde a la señal de entrada exógena y determinística. Además se tiene que $u_t = 0$ e $y_t = 0$ para $t \leq 0$.

Supuesto 5.2 (Condiciones de regularidad). El vector de parámetros asociado al sistema β_0 , la entrada u_t y el ruido ω_t en (5.1) satisfacen ciertas condiciones de regularidad, las cuales garantizan que la solución $\hat{\beta}_{ML}$ del problema de estimación mediante máxima verosimilitud converge (en probabilidad o de manera casi segura) hacia los parámetros reales del sistema β_0 a medida que $N \rightarrow \infty$.

Supuesto 5.3 (Orden del sistema). Los ordenes n_a , n_b , r , p y n_d de los polinomios del sistema (5.1) son conocidos.

Supuesto 5.4 (Distribución ruido). La distribución del ruido ω_t corresponde a una mezcla finita de distribuciones gaussianas dada por (5.2), cuyo número de componentes M es conocido.

Supuesto 5.5 (Estabilidad). El sistema (5.1) es asintóticamente estable y sus polinomios (5.4) no poseen raíces sobre el círculo unitario.

Supuesto 5.6 (Cancelaciones). El sistema real (5.1) no posee cancelaciones polo-cero, ni en su equivalente espectral de fase mínima obtenido al reflejar los ceros ubicados en el exterior del círculo unitario hacia su interior.

Supuesto 5.7 (Mediciones). El conjunto de mediciones, dado por $y_{1:N}$ es modelado como una colección de variables aleatorias.

El **Supuesto 5.2** es necesario para asegurar la identificabilidad del sistema. Esto es analizado en detalle en [75], por ende no será tratado aquí.

La identificación de los parámetros del modelo de mezcla gaussiana GMM, dados por γ , ha sido generalmente abordado añadiendo restricciones adicionales a los parámetros contenidos en el modelo [50]. No obstante, en el desarrollo del algoritmo de estimación no se consideran restricciones adicionales a la impuesta por la definición del GMM (5.2) (la suma de los pesos no negativos de la mezcla deben sumar 1).

Si bien, se asume que el sistema no se encuentra submodelado dado que se conocen los ordenes de los polinomios (ver **Supuesto 5.3**), es posible relajar dicha restricción mediante el uso de un criterio de información (ver página 43) para determinar orden correcto del sistema y el número correcto de ceros de fase mínima para la función de transferencia asociada al ruido. De manera análoga es posible relajar el **Supuesto 5.4** al utilizar un criterio que permita seleccionar el número de componentes M del GMM cuando es utilizado para aproximar la distribución desconocida del ruido ω_t (explotando el **Lema 4.3**), siguiendo una estrategia similar a las utilizadas en [50, 149].

A continuación se aborda el problema de estimación, utilizando el método de máxima verosimilitud para identificar el sistema lineal de fase no-mínima, cuando la distribución del ruido se encuentra dada (o se puede aproximar) mediante un modelo de mezcla de distribuciones gaussianas GMM.

Posteriormente se soluciona el problema mediante el algoritmo de maximización de la esperanza (EM).

Observación 5.1. *Dicha solución corresponde a una especialización de la metodología de identificación propuesta en [51] para el problema de interés.* ∇

5.2. Estimación de sistemas de fase no-mínima con mezcla de distribuciones gaussianas mediante máxima verosimilitud

En el **Capítulo 4**, particularmente en la **Sección 4.3**, se ha abordado el problema de identificación de fase mínima para el sistema dado por (5.1) (i.e. $C_u(z^{-1}, \theta) = 1$).

El error de predicción ahí presentado (4.19) es inviable en caso de que $H(z^{-1}, \theta)$ posea ceros de fase no-mínima, debido a que el predictor a un paso para y_t dado por $\hat{y}_{t|t-1}$ se vuelve inestable. El desarrollo de dicha expresión se encuentra en el **Ejemplo 2.3**.

Entonces, para poder formular el problema de estimación mediante máxima verosimilitud para (5.1), es necesario obtener en primer lugar una expresión que permita obtener el error de predicción considerando los ceros de fase mínima y no-mínima presentes en la función de transferencia $H(z^{-1}, \theta)$, cuyo desarrollo se puede apreciar a continuación.

5.2.1. Error de predicción para sistemas dinámicos de fase no-mínima

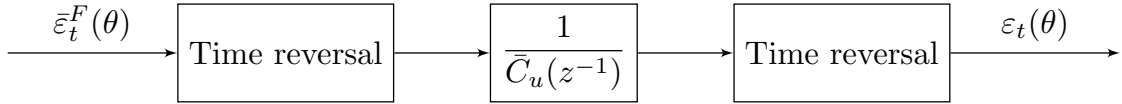


Fig. 5.1: Diagrama de bloques del esquema utilizado para realizar el filtrado anticausal y obtener $\varepsilon_t(\theta)$ en (5.10).

El polinomio que contiene los ceros de fase no-mínima (5.6), dado por $C_u(z^{-1}, \theta)$, se puede reescribir de la siguiente manera

$$\begin{aligned} C_u(z^{-1}, \theta) &= \left(1 + \frac{c_{u_{p-1}}}{c_{u_p}} z + \dots + \frac{1}{c_{u_p}} z^p \right) c_{u_p} z^{-p} \\ &= \bar{C}_u(z, \theta) c_{u_p} z^{-p}. \end{aligned} \quad (5.7)$$

El retardo de tiempo $c_{u_p} z^{-p}$ en (5.7) se puede incluir en la descripción de la secuencia de ruido ω_t en (5.1) al definir $v_t = c_{u_p} \omega_{t-p}$, con lo cual se puede redefinir (5.1) como

$$y_t = \frac{B(z^{-1}, \theta)}{A(z^{-1}, \theta)} u_t + \frac{C_s(z^{-1}, \theta) \bar{C}_u(z, \theta)}{D(z^{-1}, \theta)} v_t. \quad (5.8)$$

Entonces, el error de predicción $\varepsilon_t(\theta)$ para el sistema (5.8) se puede obtener utilizando un filtrado recursivo causal seguido por un filtrado recursivo no-causal, siguiendo la estrategia propuesta en [19, 42, 43] de la siguiente forma

$$\bar{\varepsilon}_t^F(\theta) = \left(y_t - \frac{B(z^{-1}, \theta)}{A(z^{-1}, \theta)} u_t \right) \frac{D(z^{-1}, \theta)}{C_s(z^{-1}, \theta)}. \quad (5.9)$$

para posteriormente obtener el error de predicción como

$$\varepsilon_t(\theta) = \frac{1}{\bar{C}_u(z, \theta)} \bar{\varepsilon}_t^F(\theta). \quad (5.10)$$

Note que (5.10) se puede interpretar como un **filtro no-causal estable** (ver e.g. [150, Capítulo 4]). En la Fig. 5.1 se ilustra la manera en la cual se realiza el filtrado no-causal para obtener el error de predicción en 5.10.

El polinomio $\bar{C}_u(z, \theta)$ en (5.10) corresponde a un polinomio mónico, cuyos ceros son iguales a los ceros de $C_u(z^{-1}, \theta)$ en (5.6) reflejados hacia el interior del círculo unitario.

Este esquema presentado en la Fig. 5.1 explota la propiedad de inversión temporal de la transformada- Z , la cual indica que si $X(z)$ corresponde a la transformada- Z de la secuencia x_t , entonces la transformada- Z de la secuencia de tiempo invertido x_{-t} corresponde a $X(z^{-1})$ [42, 72, 94, 150].

Observación 5.2. *Este resultado también es aplicable a sistemas de fase mínima, al considerar $C_u(z^{-1}, \theta) = 1$ en (5.8). Luego, el error de predicción se obtiene directamente a partir del filtro causal y estable dado por (5.9) [1, 2].* ∇

Observación 5.3. *La secuencia v_t corresponde a ruido blanco, cuyo espectro es igual al de ω_t escalado por c_{u_p} . Para sistemas de fase mínima, se tiene que $v_t = \omega_t$.* ∇

5.2.2. Estimación de sistemas lineales de fase no-mínima utilizando un modelo de mezcla gaussiana mediante máxima verosimilitud

A continuación, utilizando el resultado obtenido para el error de predicción obtenido previamente (ver ecuaciones (5.9) y (5.10)), es posible definir la función de verosimilitud y posteriormente formular el problema de estimación mediante máxima verosimilitud para el sistema dado por (5.1).

Para ello debemos definir en primer lugar el vector de parámetros a estimar, el cual se encuentra dado por

$$\beta = [\theta \ \gamma] \quad (5.11)$$

donde

$$\theta = [a_1 \ b_1 \ c_{s_1} \ c_{u_1} \ d_1 \ \dots \ a_{n_a} \ b_{n_b} \ c_{s_r} \ c_{u_p} \ d_{n_d}] \quad (5.12a)$$

$$\gamma = [\underbrace{\alpha_1 \ \mu_1 \ \Sigma_1}_{\gamma_1} \ \dots \ \underbrace{\alpha_M \ \mu_M \ \Sigma_M}_{\gamma_M}], \quad (5.12b)$$

donde γ corresponde al vector de parámetros del modelo de mezcla de distribuciones gaussianas GMM (5.2) para el ruido ω_t y θ representa los parámetros del modelo para el sistema (5.1). Dado que $v_t = c_{u_p} \omega_{t-p}$, se tiene que el vector de parámetros asociado a la distribución de v_t se encuentra dado por

$$\tilde{\gamma} = [\underbrace{\alpha_1 \ \tilde{\mu}_1 \ \tilde{\Sigma}_1}_{\tilde{\gamma}_1} \ \dots \ \underbrace{\alpha_M \ \tilde{\mu}_M \ \tilde{\Sigma}_M}_{\tilde{\gamma}_M}], \quad (5.13)$$

con

$$\tilde{\mu}_i = c_{u_p} \mu_i \quad (5.14)$$

$$\tilde{\Sigma}_i = c_{u_p}^2 \Sigma_i, \quad (5.15)$$

por ende, la distribución de v_t asociada al modelo (5.8) corresponde a

$$p(v_t) = \sum_{i=1}^M \alpha_i \phi(v_t; \tilde{\mu}_i, \tilde{\sigma}_i^2), \quad 0 \leq \alpha_i \leq 1, \quad \sum_{i=1}^M \alpha_i = 1, \quad (5.16)$$

con $M \geq 2$ ($M \in \mathbb{N}$) y $\phi(v_t; \tilde{\mu}_i, \tilde{\sigma}_i^2)$ representa una función de densidad de probabilidad (PDF) gaussiana con media $\tilde{\mu}_i$ y covarianza $\tilde{\Sigma}_i = \tilde{\sigma}_i^2$ dada por

$$\phi(v_t; \tilde{\mu}_i, \tilde{\sigma}_i^2) = \frac{1}{\sqrt{2\pi\tilde{\sigma}_i^2}} \exp \left\{ -\frac{(v_t - \tilde{\mu}_i)^2}{2\tilde{\sigma}_i^2} \right\}. \quad (5.17)$$

Entonces, bajo los **Supuestos 5.1 a 5.7** se tiene que la **función de verosimilitud** $\mathcal{L}(\beta)$ corresponde a

$$\begin{aligned} \mathcal{L}(\beta) &= p(y_{1:N} | \beta) \\ &= \prod_{t=1}^N \sum_{i=1}^M \alpha_i \phi(\varepsilon_t(\theta); \tilde{\mu}_i, \tilde{\Sigma}_i), \end{aligned} \quad (5.18)$$

donde $\varepsilon_t(\theta)$ corresponde al error de predicción (5.10). Por consiguiente, el **logaritmo de la función de verosimilitud** se encuentra dado por

$$\ell(\beta) = \sum_{t=1}^N \log \left\{ \sum_{i=1}^M \alpha_i \phi(\varepsilon_t(\theta); \tilde{\mu}_i, \tilde{\Sigma}_i) \right\}. \quad (5.19)$$

Luego, el problema de estimación mediante máxima verosimilitud se puede escribir como

$$\hat{\beta}_{\text{ML}} = \arg \max_{\beta} \ell(\beta) \quad \text{sujeto a} \quad 0 \leq \alpha_i \leq 1, \quad \sum_{i=1}^M \alpha_i = 1. \quad (5.20)$$

Observación 5.4 (Relación con PEM).

*Note que la función de costo dada por (5.19) corresponde al funcional de costo del método del error de predicción PEM (dado por $N^{-1} \sum_{t=1}^N \varepsilon_t^2(\theta)$), cuando el número de componentes del modelo de mezcla de distribuciones gaussianas GMM $M = 1$, la media del ruido es igual a cero $\mu_1 = 0$ y la función de transferencia $H(z^{-1}, \theta)$ es de fase mínima (ver **Ejemplo 2.2** y [1, 2]).* ▽

Al igual que en el **Capítulo 4**, la solución de (5.20) puede ser difícil de obtener, más aún cuando se incrementa el número de componentes de mezcla M en el GMM debido a que $\ell(\beta)$ presenta múltiples máximos locales [151] (ver **Observación 4.4**). Además, cada componente adicional del modelo de mezcla de distribuciones gaussianas GMM incrementa la dimensión del espacio de búsqueda en 3.

En el **Capítulo 4** se presenta una manera de solucionar el problema de optimización utilizando un algoritmo de optimización global, particularmente el algoritmo de búsqueda de patrones (Pattern Search). Si bien se obtienen resultados satisfactorios en cuanto a la covarianza del error de estimación y aproximación de la distribución del ruido, los tiempos de ejecución del algoritmo y el ajuste de los parámetros del algoritmo de búsqueda representan una dificultad.

El algoritmo de **máximización de la esperanza** [44], más conocido como Expectation-Maximization, posee ventajas respecto del algoritmo de optimización global en cuanto a tiempos de ejecución y convergencia (asumiendo una correcta inicialización del algoritmo). Además el paso-E y el paso-M son usualmente de fácil implementación. El algoritmo EM permite resolver el problema de estimación dado por (5.20) mediante una función auxiliar Q , la cual es usualmente maximizada mediante un algoritmo clásico de optimización. El algoritmo EM ha sido utilizado con anterioridad para estimar los parámetros de un modelo de mezcla de distribuciones gaussianas para sistemas estáticos [36, 50, 106, 152].

5.3. Algoritmo de maximización de la esperanza (EM) para sistemas lineales con mezcla finita de distribuciones gaussianas

En esta sección se desarrolla un algoritmo basado en la **maximización de la esperanza** (EM) para resolver el problema de identificación (5.20) de un sistema lineal de fase no-mínima cuando la distribución del ruido es aproximada mediante un modelo de mezcla de distribuciones gaussianas.

El algoritmo EM es una herramienta muy utilizada en la actualidad para la identificación de sistemas dinámicos lineales y no lineales, tanto en el dominio del tiempo (ver e.g. [46–48]) como en el dominio de la frecuencia [49].

5.3.1. Formulación del algoritmo basado en la maximización de la esperanza EM

La formulación del algoritmo EM (ver [44, 45]) con un modelo de mezcla de distribuciones gaussianas es desarrollado a partir de la definición de la función de verosimilitud con datos aumentados, siguiendo una estrategia similar a la presentada en [36, 50, 51].

Al abordar el problema de manera similar, se redefine la función de verosimilitud utilizando los datos observados $y_{1:N}$ y una variable aleatoria **latente**¹ u oculta discreta. Esta

¹El término **variable latente** es adoptado a partir de estadística y se refiere a una variable que no es directamente observada. Debido a esto, las variables latentes deben ser inferidas (mediante un modelo

variable aleatoria **latente** corresponde a un indicador que determina si una observación y_t proviene de la i -ésima componente del modelo de mezcla de distribuciones gaussianas. Luego, el algoritmo EM es desarrollado a partir de la redefinición del logaritmo de la función de verosimilitud utilizando la **información completa**² para obtener una función auxiliar (*E-step*), la cual es posteriormente maximizada respecto de los parámetros del modelo de mezcla de distribuciones gaussianas GMM (*M-step*) [45, 50].

Para el problema de interés, el logaritmo de la función de verosimilitud (5.19) es definida utilizando el error de predicción dado por (5.10). En [51] se describe un procedimiento sistemático para construir la función auxiliar de $\ell(\beta)$ (5.19). Al abordar el problema de esta manera, se obtiene un algoritmo iterativo capaz de solucionar (4.21) manteniendo las propiedades del algoritmo EM tradicional con mezcla de gaussianas. A continuación se presenta el algoritmo EM desarrollado para solucionar el problema de estimación

Teorema 5.1. *Considere el vector de parámetros a estimar dado por $\beta = [\theta \ \gamma]$ definido en (5.12). Entonces, el **paso-E** (*E-step*) del algoritmo EM se encuentra dado por*

$$\mathcal{Q}(\beta, \hat{\beta}^{(m)}) = \sum_{t=1}^N \sum_{i=1}^M \hat{\zeta}_{t,i}^{(m)} \log \left[\alpha_i \phi(\varepsilon_t(\theta); \tilde{\mu}_i, \tilde{\Sigma}_i) \right], \quad (5.21)$$

donde $\tilde{\mu}_i$ y $\tilde{\Sigma}_i$ se encuentran dados por (5.14) y (5.15) respectivamente, y

$$\hat{\zeta}_{t,i}^{(m)} = \frac{\hat{\alpha}_i^{(m)} \phi \left(\varepsilon_t(\hat{\theta}^{(m)}); \hat{\mu}_i^{(m)}, \hat{\Sigma}_i^{(m)} \right)}{\sum_{l=1}^M \hat{\alpha}_l^{(m)} \phi \left(\varepsilon_t(\hat{\theta}^{(m)}); \hat{\mu}_l^{(m)}, \hat{\Sigma}_l^{(m)} \right)}. \quad (5.22)$$

En (5.21) y (5.22) el índice m representa el número de iteración. Considere además la maximización de (5.21) respecto a β sujeto a $0 \leq \alpha_i \leq 1$ y $\sum_{i=1}^M \alpha_i = 1$. Al fijar $\hat{\theta}^{(m)}$ y optimizar (5.21) respecto a γ , se tiene que el **paso-M** del algoritmo EM se encuentra dado por

$$\hat{\alpha}_i^{(m+1)} = \frac{\sum_{t=1}^N \hat{\zeta}_{t,i}^{(m)}}{N}, \quad (5.23)$$

$$\hat{\mu}_i^{(m+1)} = \frac{\sum_{t=1}^N \hat{\zeta}_{t,i}^{(m)} \varepsilon_t(\hat{\theta}^{(m)})}{\hat{\alpha}_i^{(m)} \sum_{t=1}^N \hat{\zeta}_{t,i}^{(m)}}, \quad (5.24)$$

matemático) a partir de otras variables, las cuales si pueden ser observadas, i.e. medidas. A las variables latentes también se les conoce como **variables ocultas** o **variables no-observables** y particularmente dentro de la literatura de EM como **datos faltantes** o **datos incompletos**

²La **información completa** corresponde al conjunto de **datos observados** y los **datos ocultos** o **datos latentes**

$$\hat{\Sigma}_i^{(m+1)} = \frac{\sum_{t=1}^N \hat{\zeta}_{t,i}^{(m)} \left(\varepsilon_t(\hat{\theta}^{(m)}) - \hat{c}_{u_p}^{(m)} \hat{\mu}_i^{(m)} \right)^2}{\left[\hat{c}_{u_p}^{(m)} \right]^2 \sum_{t=1}^N \hat{\zeta}_{t,i}^{(m)}}. \quad (5.25)$$

Luego, para optimizar (5.21) respecto de θ se debe solucionar

$$\hat{\theta}^{(m+1)} = \arg \min_{\theta} \mathcal{B}(\theta), \quad (5.26)$$

donde

$$\mathcal{B}(\theta) = \sum_{t=1}^N \sum_{i=1}^M \frac{\hat{\zeta}_{t,i}^{(m)}}{\left[\hat{c}_{u_p}^{(m)} \right]^2 \hat{\Sigma}_i^{(m+1)}} \left(\varepsilon_t(\theta) - \hat{c}_{u_p}^{(m)} \hat{\mu}_i^{(m+1)} \right)^2. \quad (5.27)$$

Observación 5.5. La optimización de la función auxiliar $\mathcal{Q}(\beta, \hat{\beta}^{(m)})$ dada por (5.21) respecto de β se realiza mediante el algoritmo de descenso por coordenadas, más conocido por su nombre en inglés *coordinate descent algorithm* [153]. Para ello, en primer lugar, se fija el valor del parámetro θ de la iteración actual dado por $\hat{\theta}^{(m)}$ y se optimiza (5.21) respecto de los parámetros asociados al modelo de mezcla de distribuciones gaussianas γ . Posteriormente se fijan los parámetros del GMM en su valor siguiente, es decir $\hat{\gamma}^{(m+1)}$ y se soluciona el problema de optimización dado por (5.26). ∇

Demostración 5.1 (Obtención de la función auxiliar $\mathcal{Q}(\beta, \hat{\beta}^{(m)})$).

En primer lugar se define

$$K(y_t, \beta_i) = \alpha_i \phi(\varepsilon_t(\theta); \tilde{\mu}_i, \tilde{\Sigma}_i), \quad (5.28)$$

$$\mathcal{V}_t(\beta) = \sum_{i=1}^M K(y_t, \beta_i), \quad (5.29)$$

donde $\beta_i = [\theta \ \gamma_i]$. Entonces, el logaritmo de la función de verosimilitud (5.19) se puede expresar como

$$\ell(\beta) = \sum_{t=1}^N \log [\mathcal{V}_t(\beta)]. \quad (5.30)$$

Al definir $\mathcal{B}_t(\beta) = \log [\mathcal{V}_t(\beta)]$, se obtiene

$$\mathcal{B}_t(\beta) = \bar{\mathcal{Q}}_t(\beta, \hat{\beta}^{(m)}) - \bar{\mathcal{H}}_t(\beta, \hat{\beta}^{(m)}), \quad (5.31)$$

donde las funciones $\bar{\mathcal{Q}}_t(\beta, \hat{\beta}^{(m)})$ y $\bar{\mathcal{H}}_t(\beta, \hat{\beta}^{(m)})$ se encuentran dadas por

$$\bar{\mathcal{Q}}_t(\beta, \hat{\beta}^{(m)}) = \sum_{i=1}^M \log [K(y_t, \beta_i)] \frac{K(y_t, \hat{\beta}_i^{(m)})}{\mathcal{V}_t(\hat{\beta}^{(m)})}, \quad (5.32)$$

$$\bar{\mathcal{H}}_t(\beta, \hat{\beta}^{(m)}) = \sum_{i=1}^M \log \left[\frac{K(y_t, \beta_i)}{\mathcal{V}_t(\beta)} \right] \frac{K(y_t, \hat{\beta}_i^{(m)})}{\mathcal{V}_t(\hat{\beta}^{(m)})}. \quad (5.33)$$

La función $\bar{\mathcal{H}}_t(\beta, \hat{\beta}^{(m)})$ es decreciente para todo valor de β . Este resultado es consecuencia de la **desigualdad de Jensen** dada por la **Definición A.6** (ver e.g. [51] y sus referencias). Entonces, el logaritmo de la función de verosimilitud, dado por (5.30), satisface lo siguiente

$$\ell(\hat{\beta}^{(m+1)}) \geq \ell(\hat{\beta}^{(m)}). \quad (5.34)$$

Al sustituir (5.31) en (5.30) y utilizar (5.28) y (5.29) se obtiene la función auxiliar $\mathcal{Q}(\beta, \hat{\beta}^{(m)})$ como

$$\mathcal{Q}(\beta, \hat{\beta}^{(m)}) = \sum_{t=1}^N \sum_{i=1}^M \hat{\zeta}_{t,i}^{(m)} \log \left[\alpha_i \phi(\varepsilon_t(\theta); \tilde{\mu}_i, \tilde{\Sigma}_i) \right], \quad (5.35)$$

donde

$$\hat{\zeta}_{t,i}^{(m)} = \frac{\hat{\alpha}_i^{(m)} \phi(\varepsilon_t(\hat{\theta}^{(m)}); \hat{\mu}_i^{(m)}, \hat{\Sigma}_i^{(m)})}{\sum_{l=1}^M \hat{\alpha}_l^{(m)} \phi(\varepsilon_t(\hat{\theta}^{(m)}); \hat{\mu}_l^{(m)}, \hat{\Sigma}_l^{(m)})}. \quad (5.36)$$

□

Demostración 5.2 (Optimización de la función auxiliar $\mathcal{Q}(\beta, \hat{\beta}^{(m)})$).

Al derivar (5.35) respecto de $\tilde{\mu}_i$ e igualar el resultado a cero se obtiene

$$\frac{\partial \mathcal{Q}(\beta, \hat{\beta}^{(m)})}{\partial \tilde{\mu}_i} = \sum_{t=1}^N \left(\varepsilon_t(\hat{\theta}^{(m)}) - \hat{\mu}_i^{(m+1)} \right) \hat{\zeta}_{t,i}^{(m)} = 0 \quad (5.37)$$

Al reemplazar (5.14) se obtiene

$$\sum_{t=1}^N \hat{\zeta}_{t,i}^{(m)} \varepsilon_t(\hat{\theta}^{(m)}) = \hat{c}_{u_p}^{(m)} \hat{\mu}_i^{(m+1)} \sum_{t=1}^N \hat{\zeta}_{t,i}^{(m)}, \quad (5.38)$$

$$\hat{\mu}_i^{(m+1)} = \frac{\sum_{t=1}^N \hat{\zeta}_{t,i}^{(m)} \varepsilon_t(\hat{\theta}^{(m)})}{\hat{c}_{u_p}^{(m)} \sum_{t=1}^N \hat{\zeta}_{t,i}^{(m)}}. \quad (5.39)$$

Siguiendo el mismo procedimiento, considere ahora la derivada de (5.35) respecto de $\lambda = \tilde{\Sigma}_i^{-1}$. Al igualar dicho resultado a cero se tiene

$$\frac{\partial \mathcal{Q}(\beta, \hat{\beta}^{(m)})}{\partial \lambda} = \sum_{t=1}^N \left[\frac{1}{\lambda} - \left(\varepsilon_t(\hat{\theta}^{(m)}) - \hat{\mu}_i^{(m)} \right)^2 \right] \hat{\zeta}_{t,i}^{(m)} = 0. \quad (5.40)$$

Lo cual se puede reescribir como

$$\sum_{t=1}^N \hat{\zeta}_{t,i}^{(m)} \left(\varepsilon_t(\hat{\theta}^{(m)}) - \hat{\mu}_i^{(m)} \right)^2 = \hat{\Sigma}_i^{(m+1)} \sum_{t=1}^N \hat{\zeta}_{t,i}^{(m)}. \quad (5.41)$$

Al reemplazar (5.14) y (5.15) en (5.41) se obtiene

$$\hat{\Sigma}_i^{(m+1)} = \frac{\sum_{t=1}^N \hat{\zeta}_{t,i}^{(m)} (\varepsilon_t(\hat{\theta}^{(m)}) - \hat{c}_{u_p}^{(m)} \hat{\mu}_i^{(m)})^2}{\left[\hat{c}_{u_p}^{(m)}\right]^2 \sum_{t=1}^N \hat{\zeta}_{t,i}^{(m)}}. \quad (5.42)$$

Para el parámetro α_i se define $\mathcal{R}(\alpha_i)$ de la siguiente forma

$$\mathcal{R}(\alpha_i) = \sum_{t=1}^N \sum_{i=1}^M \log[\alpha_i] \hat{\zeta}_{t,i}^{(m)}, \quad (5.43)$$

sujeto a $\sum_{i=1}^M \alpha_i = 1$. Note que no se ha considerado la restricción $0 \leq \alpha_i \leq 1$ en la definición. Entonces, utilizando un **multiplicador de Lagrange** [154, Capítulo 5] para considerar la restricción sobre α_i , se define

$$\mathcal{G}(\alpha_i, \rho) = \sum_{t=1}^N \sum_{i=1}^M \log[\alpha_i] \hat{\zeta}_{t,i}^{(m)} - \rho \left(\sum_{i=1}^M \alpha_i - 1 \right). \quad (5.44)$$

Al derivar (5.44) respecto de α_i y ρ e igualando los resultados a cero se obtiene

$$\frac{\partial \mathcal{G}(\alpha_i, \rho)}{\partial \alpha_i} = \frac{\sum_{t=1}^N \hat{\zeta}_{t,i}^{(m)}}{\hat{\alpha}_i^{(m+1)}} - \rho = 0. \quad (5.45)$$

$$\frac{\partial \mathcal{G}(\alpha_i, \rho)}{\partial \rho} = \sum_{i=1}^M \alpha_i - 1 = 0. \quad (5.46)$$

Por lo cual

$$\hat{\alpha}_i^{(m+1)} = \frac{\sum_{t=1}^N \hat{\zeta}_{t,i}^{(m)}}{\rho}. \quad (5.47)$$

Al sumar (5.47) de $i = 1 \dots M$ y reemplazando (5.46) se tiene

$$\sum_{i=1}^M \hat{\alpha}_i^{(m+1)} = \frac{\sum_{i=1}^M \sum_{t=1}^N \hat{\zeta}_{t,i}^{(m)}}{\rho} = 1, \quad (5.48)$$

$$\rho = \sum_{i=1}^M \sum_{t=1}^N \hat{\zeta}_{t,i}^{(m)}. \quad (5.49)$$

Note que $\sum_{i=1}^M \hat{\zeta}_{t,i}^{(m)} = 1$, por ende $\rho = N$. Finalmente se obtiene

$$\hat{\alpha}_i^{(m+1)} = \frac{\sum_{t=1}^N \hat{\zeta}_{t,i}^{(m)}}{N}. \quad (5.50)$$

Note que la restricción $0 \leq \hat{\alpha}_i^{(m+1)} \leq 1$ se cumple, pese a no ser explícitamente considerada en (5.44). $\square$

Finalmente, es posible resumir el algoritmo de la siguiente manera:

Paso-I Fije el número de componentes de la mezcla de distribuciones gaussianas M a utilizar.

Paso-II Seleccione una estimación inicial para $\hat{\beta}^{(0)}$ y fije el número de iteración en cero $m = 0$.

Paso-III Calcule el error de predicción $\varepsilon_t(\hat{\theta}^{(m)})$ a partir de (5.9) y (5.10).

Paso-IV Estime los parámetros asociados al modelo de mezcla de distribuciones gaussianas que aproximan la distribución de ω_t utilizando las ecuaciones (5.23) a (5.25).

Paso-V Estime los parámetros del sistema mediante (5.26).

Paso-VI Establezca $m = m + 1$ y vuelva al **Paso-III** hasta que se cumpla el criterio de detención del algoritmo.

5.3.2. Procedimiento de inicialización para el algoritmo EM con GMM

En [155–157] se muestra que la precisión de la estimación de los algoritmos EM con GMM depende en gran medida de la estimación inicial de los parámetros de la mezcla gaussiana $\hat{\gamma}^{(0)}$.

Diversas técnicas de inicialización para los componentes de la mezcla gaussiana han sido propuestos y estudiados en los últimos años (ver e.g. [151, 155, 158, 159] entre otros). En dichos trabajos se muestra que una inicialización adecuada y meticulosa de los parámetros del modelo de mezcla gaussiana GMM para el algoritmo EM es fundamental para obtener estimaciones precisas y mejorar la convergencia.

Para el problema dado por (5.1) se utiliza el método del error de predicción PEM (ver **Sección 2.4** y [1, 2]) para obtener una estimación inicial de los parámetros del sistema $\hat{\theta}^{(0)}$. Posteriormente se obtiene el error de predicción, mediante (5.9) y (5.10), el cual proporciona las estadísticas de la distribución del ruido necesarias para inicializar los parámetros del GMM $\hat{\gamma}^{(0)}$.

Con la finalidad de ilustrar este procedimiento, considere un sistema (MA) simple (ver Tabla 2.1) para (5.1), con $C_s(z^{-1}, \theta) = 1$ y $C_u(z^{-1}, \theta) = 1 - 1,5z^{-1}$. Considere adicionalmente que la distribución de ω_t corresponde a un GMM con $M = 2$ componentes (5.2), cuyos pesos se encuentran dados por $\alpha_1 = \alpha_2 = 0,5$, varianzas $\Sigma_1 = \Sigma_2 = 1$ y medias $\mu_1 = -2$ y $\mu_2 = 2$. Entonces, los parámetros asociados a la distribución v_t del modelo (5.8) se encuentran dados por $\tilde{\Sigma}_1 = \tilde{\Sigma}_2 = 2,25$, $\tilde{\mu}_1 = -3$ y $\tilde{\mu}_2 = 3$. Luego, el

método del error de predicción PEM es utilizado para estimar los parámetros iniciales del modelo $\hat{\theta}^{(0)} = \hat{\theta}^{(\text{PEM})}$. Posteriormente, utilizando (5.9) y (5.10) se obtiene el error de predicción $\varepsilon_t(\hat{\theta}^{(0)})$. En la Fig. 5.2 se puede apreciar el error de predicción obtenido. A partir de la información proporcionada por el error de predicción (ver Fig. 5.2) se obtiene una inicialización para los parámetros del GMM $\hat{\gamma}^{(0)}$.

La inicialización del algoritmo propuesto (ver **Paso-II**) se puede resumir de la siguiente manera:

- (1) El valor inicial para los parámetros asociados al modelo del sistema, $\hat{\theta}^{(0)}$, se obtienen utilizando el método del error de predicción PEM.
- (2) Los valores medios iniciales para los componentes del GMM, dados por $\hat{\mu}_i^{(0)}$, se eligen equiespaciados entre el valor mínimo y máximo para el error de predicción $\varepsilon_t(\hat{\theta}^{(0)})$.
- (3) Las varianzas iniciales $\{\hat{\Sigma}_i^{(0)}\}_{i=1}^M$ del GMM se establecen iguales a la varianza muestral del error de predicción $\varepsilon_t(\hat{\theta}^{(0)})$.
- (4) Los pesos iniciales $\hat{\alpha}_i^{(0)}$ de los componentes del GMM se eligen iguales de modo que $\sum_{i=1}^M \hat{\alpha}_i^{(0)} = 1$.

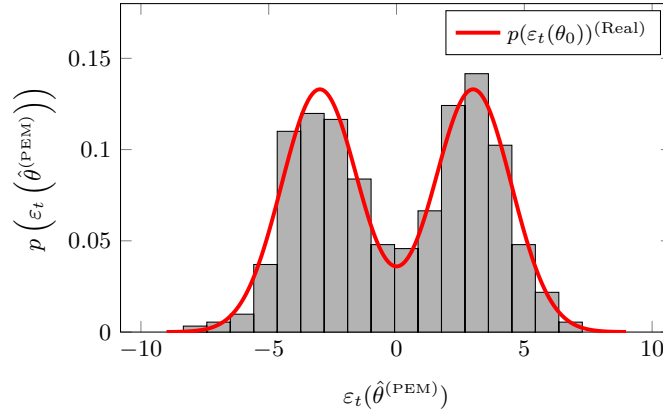


Fig. 5.2: Histograma de $\varepsilon_t(\hat{\theta}^{(0)})$ obtenido a partir de (5.9) y (5.10), con $\hat{\theta}^{(0)} = \hat{\theta}^{(\text{PEM})}$.

5.4. Ejemplos numéricos

A continuación se presentan 3 ejemplos numéricos con la finalidad de analizar el desempeño del algoritmo EM propuesto (ver **Teorema 5.1**) para resolver el problema de estimación mediante máxima verosimilitud de un sistema dinámico, el cual puede ser de fase no-mínima, cuando la distribución del ruido es no gaussiana. Considere entonces,

que las observaciones son generadas a partir del siguiente sistema

$$y_t = \frac{b_1^o z^{-1}}{1 + a_1^o z^{-1}} u_t + \frac{1 + c_1^o z^{-1}}{1 + d_1^o z^{-1}} \omega_t, \quad (5.51)$$

cuyos parámetros para el *sistema real* se encuentran dados por $b_1^o = 1$, $a_1^o = -0,5$, $d_1^o = 0,8$. La señal de entrada se encuentra dada por $u_t \sim \mathcal{N}(0, \sigma_u^2)$ con $\sigma_u^2 = 10$, mientras que el ruido ω_t posee una distribución no-gaussiana.

Nos centraremos en dos escenarios para (5.51) considerando una función de transferencia de fase mínima para $H(z^{-1}, \theta)$ con $c_1 = -0,1$. Luego se considera el problema de fase no-mínima para la función de transferencia asociada al ruido, con $c_1^o = 1,25$. El vector de parámetros a estimar se encuentra dado por $\beta = [\theta \ \gamma]$, con $\theta = \{b_1, a_1, c_1, d_1\}$ y $\gamma = \{\alpha_i, \mu_i, \Sigma_i\}_{i=1}^M$.

Para el primer ejemplo se considera un sistema de fase no-mínima para $H(z^{-1}, \theta)$ en (5.51) considerando los **Supuestos 5.1 a 5.7**. La distribución del ruido ω_t se encuentra dada por una mezcla de distribuciones gaussianas, cuyo número de componentes M es conocido.

Posteriormente se consideran 2 ejemplos en los cuales se relaja el **Supuesto 5.4** al considerar distribuciones no-gaussianas para ω_t . No obstante, dichas distribuciones pueden ser aproximadas mediante un modelo de mezcla de distribuciones gaussianas, lo cual fue abordado previamente en el **Capítulo 4**, particularmente en el **Lema 4.3**.

5.4.1. Configuración de las simulaciones

La configuración de la simulación utilizada en los ejemplos es la siguiente:

- (1) La inicialización de los parámetros del algoritmo $\hat{\beta}^{(0)}$ se realiza mediante el procedimiento expuesto en la **Sección 5.3.2**.
- (2) Se considera una longitud de datos dada por $N = 5000$.
- (3) El número de simulaciones de Monte-Carlo (MC) es de 100.
- (4) El criterio de detención se encuentra dado por

$$\left\| \hat{\beta}^{(m)} - \hat{\beta}^{(m-1)} \right\| / \left\| \hat{\beta}^{(m)} \right\| < 10^{-6},$$

donde m corresponde al número de iteración, o cuando se logra el número máximo de iteraciones dado por $m_{max} = 2000$.

5.4.2. Ejemplo 1: Sistema de fase no-mínima con distribución de ruido dada por una mezcla gaussiana superpuesta

Considere el sistema definido por (5.51), con ω_t distribuido como mezcla de distribuciones gaussianas (5.2) con $M = 3$. Los parámetros de la mezcla se encuentran dados

por $\mu_1 = -3$, $\mu_2 = 0$, $\mu_3 = 3$ para las medias, $\Sigma_1 = \Sigma_2 = \Sigma_3 = 1$ para las varianzas y $\alpha_1 = \alpha_3 = 0,25$, $\alpha_2 = 0,5$ para los pesos de la mezcla.

En la Fig. 5.4 se puede apreciar el promedio de las distribuciones estimadas en la simulación de Monte-Carlo, dado por $\hat{p}(w_t)^{(GMM)}$. Al igual que en el **Capítulo 4**, el área sombreada de color gris presente en la figura representa la región de incertidumbre (RI), i.e. el área donde se encuentran todas las aproximaciones de la función de densidad de probabilidad PDF para la simulación de Monte-Carlo. Se puede observar que el modelo de mezcla de distribuciones gaussianas GMM estimado se ajusta a la verdadera distribución del ruido, dada por $\hat{p}(w_t)^{(Real)}$.

Adicionalmente, se estiman los parámetros del modelo utilizando la metodología para los sistemas de fase no-mínima propuesta en el **Capítulo 3**, la cual se encuentra basada en los momentos de alto orden (MoM) [16]. En la Fig. 5.3 se presentan los resultados de la estimación paramétrica para el sistema de fase no-mínima (5.51). En dichas figuras, la cruz roja representa el valor real de los parámetros del sistema, i.e. θ_0 . Note que la estimación obtenida utilizando el método propuesto, el cual denominaremos EM-GMM, presenta una menor dispersión respecto del método de los momentos (MoM). En la Tabla 5.1 se presentan los resultados estadísticos, de la estimación paramétrica del sistema para ambos métodos utilizados (EM-GMM y MoM), en la simulación de Monte-Carlo. Se puede ver como el método propuesto presenta una mejor precisión en la estimación de los parámetros del modelo θ respecto del método de los momentos MoM.

Método	$\hat{a}_1$	$\hat{b}_1$	$\hat{c}_1$	$\hat{d}_1$
MoM	$-0,5009 \pm 1,4 \times 10^{-2}$	$0,9992 \pm 1,7 \times 10^{-2}$	$1,5081 \pm 8,6 \times 10^{-2}$	$0,7994 \pm 2,9 \times 10^{-2}$
EM-GMM	$-0,5001 \pm 5,4 \times 10^{-3}$	$1,0001 \pm 8,5 \times 10^{-3}$	$1,5005 \pm 1,1 \times 10^{-2}$	$0,8002 \pm 3,6 \times 10^{-3}$

Tabla 5.1: Estadísticas de la simulación de Monte-Carlo para el Ejemplo 1.

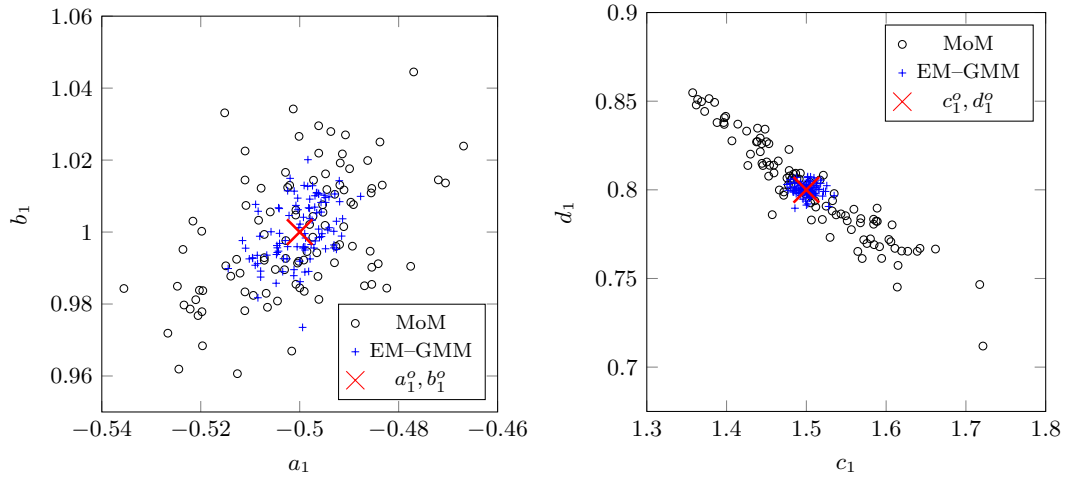
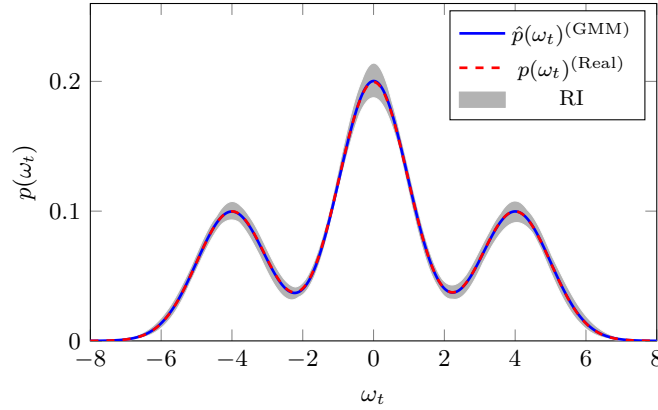


Fig. 5.3: Resultados de la estimación paramétrica para el Ejemplo 1.

Fig. 5.4: Aproximación de la distribución del ruido ω_t al utilizar un GMM para el Ejemplo 1.

5.4.3. Ejemplo 2: Sistema de fase no-mínima con ruido de distribución uniforme

Considere el sistema definido por (5.51), con ω_t distribuido de manera uniforme en el intervalo $[-2, 2]$, i.e. $\omega_t \sim \mathcal{U}[-2, 2]$. Para poder aproximar la distribución del ruido mediante un GMM, se consideran $M = 7$ componentes de mezcla.

En la Fig. 5.6 es posible observar el promedio de las distribuciones estimadas en la simulación de Monte-Carlo, dado por $\hat{p}(w_t)^{(GMM)}$. Al igual que en el ejemplo anterior, el área sombreada de color gris corresponde a la región de incertidumbre. Se puede ver que el GMM con $M = 7$ componentes se ajusta adecuadamente a la verdadera distribución del ruido, dada por $\hat{p}(w_t)^{(Real)}$.

Con la finalidad de comparar los resultados obtenidos, se estiman los parámetros del sistema mediante el método de los momentos (MoM) (descrito en la **Sección 3.4**). En la Fig. 5.5 se pueden apreciar los resultados de la estimación paramétrica para el sistema de fase no-mínima (5.51) cuando el ruido ω_t se encuentra distribuido de manera uniforme. La cruz roja presente en dichas figuras corresponde al valor real de los parámetros del sistema θ_0 . Se puede apreciar que el desempeño del método EM-GMM para estimar los parámetros del sistema es considerablemente mejor que el método de los momentos (MoM), en cuanto a la precisión de la estimación. Esta afirmación es confirmada por los resultados estadísticos de la simulación de Monte-Carlo para el presente ejemplo expuestos en la Tabla 5.2, en la cual se puede apreciar que el método propuesto mejora la precisión de la estimación respecto del método de los momentos.

Método	$\hat{a}_1$	$\hat{b}_1$	$\hat{c}_1$	$\hat{d}_1$
MoM	$-0,5007 \pm 4,9 \times 10^{-3}$	$0,9990 \pm 7,3 \times 10^{-3}$	$1,5114 \pm 8,6 \times 10^{-2}$	$0,7982 \pm 2,9 \times 10^{-2}$
EM-GMM	$-0,4998 \pm 1,3 \times 10^{-3}$	$0,9999 \pm 1,7 \times 10^{-3}$	$1,5003 \pm 5,9 \times 10^{-3}$	$0,8001 \pm 2,3 \times 10^{-3}$

Tabla 5.2: Estadísticas de la simulación de Monte-Carlo para el Ejemplo 2.

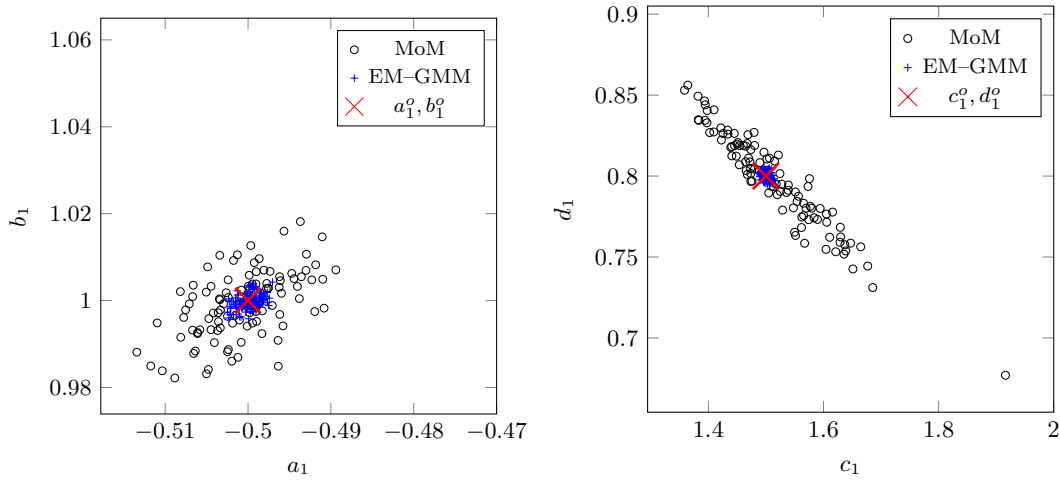


Fig. 5.5: Resultados de la estimación paramétrica para el Ejemplo 2.

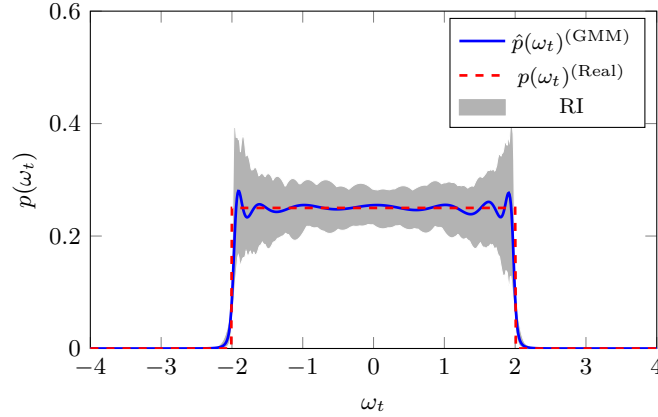


Fig. 5.6: Aproximación de la distribución del ruido ω_t al utilizar un GMM para el Ejemplo 2.

5.4.4. Ejemplo 3: Sistema de fase mínima con ruido de distribución Chi-cuadrado

En este ejemplo se considera el sistema (5.51) de fase mínima, cuyo ruido ω_t posee una distribución Chi-cuadrado, $\omega_t \sim \chi^2(\kappa)$ con $\kappa = 2$. Dado que este sistema es de fase mínima, es posible obtener el error de predicción directamente a partir de (5.9) $\varepsilon_t(\theta) = \bar{\varepsilon}_t^F(\theta)$ (ver **Observación 5.2**). Al igual que en el ejemplo anterior, se consideran $M = 7$ componentes de mezcla para aproximar la distribución del ruido mediante un GMM.

En la Fig. 5.8 se presenta el promedio de las distribuciones estimadas en la simulación de Monte-Carlo, dado por $\hat{p}(w_t)^{(GMM)}$. Al igual que en los ejemplos anteriores, el área sombreada de color gris corresponde a la región de incertidumbre. Se puede apreciar claramente, que el GMM con $M = 7$ componentes se ajusta adecuadamente a la verdadera distribución del ruido, dada por $\hat{p}(w_t)^{(Real)}$.

Para contrastar los resultados obtenidos se utiliza el método del error de predicción PEM [1] (ver **Sección 2.4**). Los resultados de la estimación utilizando el método propuesto y PEM se pueden apreciar en la Fig. 5.7. En ambas figuras, la cruz roja representa los valores verdaderos del sistema, es decir θ_0 . Se puede apreciar que la estimación obtenida con el método propuesto es considerablemente más precisa que el método del error de predicción PEM. Además se observa un sesgo en los resultados obtenidos mediante PEM, particularmente en la identificación de los parámetros asociados a la función de transferencia del ruido.

En la Tabla 5.3 se presenta la media de las estimaciones y su correspondiente desviación estándar para ambos algoritmos, i.e. EM-GMM y PEM. Se puede apreciar que mediante el método propuesto se obtiene una estimación precisa de los parámetros del

sistema, lo cual ilustra el beneficio de considerar un modelo de mezcla de distribuciones gaussianas para aproximar la distribución del ruido cuando es no-gaussiana.

Método	$\hat{a}_1$	$\hat{b}_1$	$\hat{c}_1$	$\hat{d}_1$
PEM	$-0,5089 \pm 4,0 \times 10^{-3}$	$1,0058 \pm 6,9 \times 10^{-3}$	$0,4810 \pm 1,4 \times 10^{-2}$	$0,8672 \pm 8,5 \times 10^{-3}$
EM-GMM	$-0,4999 \pm 6,0 \times 10^{-4}$	$1,0001 \pm 9,0 \times 10^{-4}$	$0,1000 \pm 2,4 \times 10^{-3}$	$0,8001 \pm 1,6 \times 10^{-3}$

Tabla 5.3: Estadísticas de la simulación de Monte-Carlo para el Ejemplo 3.

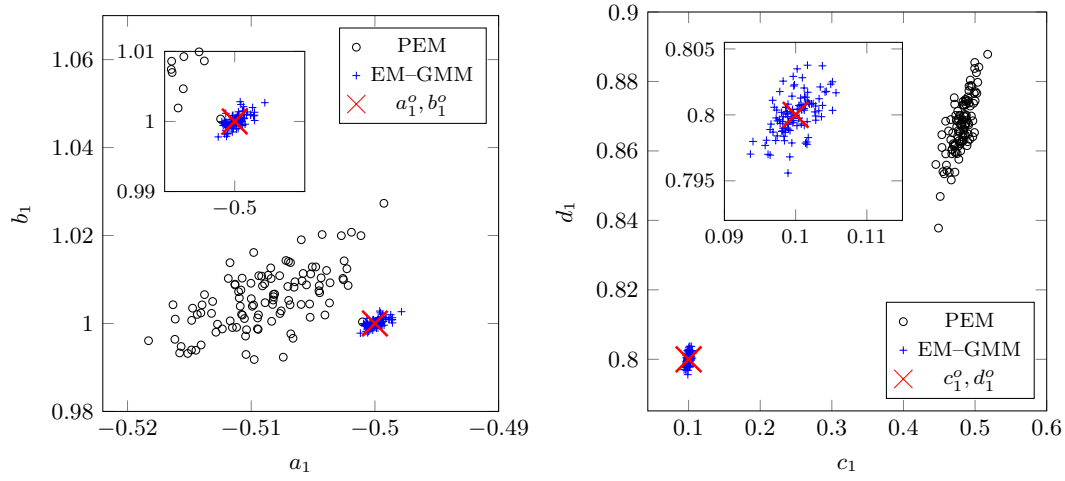


Fig. 5.7: Resultados de la estimación paramétrica para el Ejemplo 3.

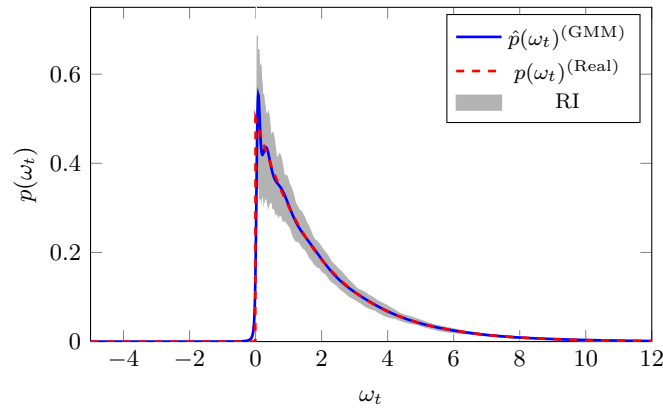


Fig. 5.8: Aproximación de la distribución del ruido ω_t al utilizar un GMM para el Ejemplo 3.

CONCLUSIONES

6.1. Resumen de los capítulos de la tesis destacando las contribuciones

A continuación se resumen las contribuciones principales de cada capítulo:

- El **Capítulo 1** introduce el problema de interés y los problemas suscitados al abordarlo desde una perspectiva clásica. Además presenta el estado del arte asociado al problema y la notación utilizada.
- El **Capítulo 2** presenta una introducción a la identificación de sistemas. En él se detallan los elementos principales de la identificación de sistemas y posteriormente se describen 2 métodos de identificación, dados por el método de máxima verosimilitud y el método del error de predicción. Finalmente se presentan algunas propiedades y definiciones relevantes.
- En el **Capítulo 3** se describe el método de los momentos, el cual representa una solución al problema de interés. Adicionalmente se exponen algunas propiedades y definiciones relevantes, junto con analizar las ventajas y desventajas del método.
- El **Capítulo 4** se centra en el desarrollo y análisis de una solución al problema de interés, restringido al caso de fase mínima para las dinámicas del ruido. Los beneficios del método se ilustran mediante simulaciones numéricas, las cuales son contrastadas con el método del error de predicción.
- El **Capítulo 5** presenta una extensión de los resultados del **Capítulo 4** al considerar ceros de fase no-mínima en las dinámicas del ruido y resolver el problema de identificación mediante el algoritmo de maximización de la esperanza EM. El desempeño del método es analizado mediante ejemplos numéricos, contrastando los resultados obtenidos con el método del error de predicción y el método de los momentos para los casos de fase mínima y no-mínima respectivamente.

6.2. Conclusiones

En este trabajo se analiza el impacto en la precisión de la estimación de sistemas dinámicos lineales invariantes en el tiempo, al considerar que la distribución del ruido es no-gaussiana y desconocida. Para ello se explota el hecho de que cualquier distribución no-gaussiana puede ser aproximada de manera arbitrariamente cercana mediante un modelo de mezcla de distribuciones gaussianas, denominado GMM (ver **Lema 4.3**).

Con la finalidad de evaluar el efecto sobre la precisión de la estimación, se analiza el problema más simple dado por la estimación de una constante en presencia de ruido aditivo no-gaussiano. Se muestra que el estimador de máxima verosimilitud presenta mejores resultados, al reducir la covarianza del error de estimación, comparado con algoritmos clásicos que no consideran distribuciones no-gaussianas en su desarrollo. No obstante, la función de verosimilitud presenta máximos locales cuando se utiliza un modelo de mezcla de distribuciones gaussianas para aproximar ruidos con distribuciones no-gaussianas.

Posteriormente se desarrolla un estimador para sistemas lineales invariantes en el tiempo de fase mínima (para la función de transferencia $H(q^{-1}, \theta)$ asociada a las dinámicas del ruido), cuando la distribución del ruido es no-gaussiana. Para solucionar el problema de optimización asociado se utiliza un algoritmo de optimización global debido a la naturaleza multimodal de la función de verosimilitud.

A pesar que la carga computacional que involucra el método propuesto es considerablemente mayor que para el método del error de predicción, se obtienen mejores resultados al identificar el sistema considerando un GMM para aproximar la distribución del ruido, en términos de la covarianza del error de estimación $P_{\hat{\theta}}$ y la norma ℓ_1 del error de estimación de la función de densidad de probabilidad del ruido $\epsilon(\eta)$.

Más adelante, en el **Capítulo 5** se presenta una extensión del estimador desarrollado, al admitir la existencia de ceros de fase no-mínima en la transferencia asociada al ruido. Para ello es necesario extender la expresión del error de predicción, considerando los ceros de fase no-mínima. Luego utilizando la nueva expresión para el error de predicción se define el problema de estimación mediante máxima verosimilitud para estimar simultáneamente los parámetros del sistema y del modelo de mezcla de distribuciones gaussianas que aproxima la distribución del ruido. Utilizando el algoritmo de maximización de la esperanza EM y el algoritmo de optimización de descenso por coordenadas, se obtiene una solución con expresiones cerradas para el problema de estimación. Mediante simulaciones numéricas se analiza el desempeño del algoritmo propuesto al ser contrastado con el método del error de predicción y el método de los momentos para fase mínima y fase no-mínima respectivamente. El algoritmo propuesto presenta mejoras en cuanto a la precisión de la estimación, especialmente en la transferencia asociada al ruido. Adicionalmente se obtiene una aproximación bastante precisa de la distribución

del ruido.

6.3. Trabajos futuros

A continuación se proponen algunos posibles trabajos de investigación futuros, los cuales extenderían los resultados obtenidos

- Utilizar el método propuesto sobre otros conjuntos de modelos, como variables de estado tanto lineal como no-lineal, error de salida no-lineal, sistemas lineales variantes en el tiempo entre otros (ver e.g. [1, Capítulos 4 y 5]).
- Extender el método propuesto a sistemas inestables ($G(q^{-1}, \theta)$ en (1.1)) utilizando la estrategia propuesta en [43].
- Analizar el comportamiento del método propuesto cuando el sistema se encuentra **submodelado** y **sobremodelado** por el conjunto de modelos $\mathcal{M}(\theta)$. Además incluir algún criterio de información (o test de hipótesis) que permita seleccionar adecuadamente el orden del modelo (es decir, los órdenes n_a , n_b , r , p y n_d de los polinomios presentes en (5.1)), como por ejemplo el criterio de información de Akaike (AIC) y su versión modificada (MAIC) (ver (3.42)), la longitud de descripción mínima (MDL), entre otros [1, Capítulo 16] [2, Capítulo 11].
- Utilizar una estrategia similar a la presentada en [149] para seleccionar un número adecuado de componentes gaussianas M del modelo de mezcla GMM.
- Extender el método a sistemas multivariados (MIMO) y contrastar su desempeño con los algoritmos de identificación comúnmente utilizados.
- Estimar la condición inicial del filtrado en tiempo reverso utilizado en (5.10) cuando la función de transferencia asociada al ruido posee ceros de fase no-mínima.
- Incluir las condiciones iniciales del sistema en la función de verosimilitud (4.21) y (5.20) siguiendo un procedimiento similar al presentado en [160, 161].

REFERENCIAS

- [1] L. Ljung, Ed., *System Identification: Theory for the User*. Upper Saddle River, NJ, USA: Prentice Hall PTR, 1999.
- [2] T. Söderström and P. Stoica, Eds., *System Identification*. Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 1988.
- [3] G. C. Goodwin and R. L. Payne, *Dynamic system identification : experiment design and data analysis*. Academic Press New York, 1977.
- [4] L. Ljung, G. C. Goodwin, and J. C. Agüero, “Stochastic Embedding revisited: A modern interpretation,” in *53rd IEEE Conference on Decision and Control*, Dec 2014, pp. 3340–3345.
- [5] A. Moschitta, J. Schoukens, and P. Carbone, “Parametric System Identification Using Quantized Data,” *IEEE Transactions on Instrumentation and Measurement*, vol. 64, no. 8, pp. 2312–2322, 2015.
- [6] B. I. Godoy, J. C. Agüero, R. Carvajal, G. C. Goodwin, and J. I. Yuz, “Identification of sparse FIR systems using a general quantisation scheme,” *Int. J. Control*, vol. 87, no. 4, pp. 874–886, 2014.
- [7] J. C. Agüero, K. González, and R. Carvajal, “EM-based identification of ARX systems having quantized output data,” *IFAC-PapersOnLine*, vol. 50, no. 1, pp. 8367 – 8372, 2017, 20th IFAC World Congress.
- [8] M. Zamani, B. Ninness, and J. C. Agüero, “On identification of Networked Systems with Time-invariant Topology,” *IFAC-PapersOnLine*, vol. 48, no. 28, pp. 1184 – 1189, 2015, 17th IFAC Symposium on System Identification.
- [9] F. Chen, J. C. Agüero, M. Gilson, H. Garnier, and T. Liu, “EM-based identification of continuous-time ARMA Models from irregularly sampled data,” *Automatica*, vol. 77, pp. 293 – 301, 2017.

- [10] D. B. Levey and S. McLaughlin, "The statistical nature of impulse noise interarrival times in digital subscriber loop systems," *Signal Processing*, vol. 82, no. 3, pp. 329 – 351, 2002.
- [11] W. Li, Y. Jia, J. Du, and J. Zhang, "PHD filter for multi-target tracking with glint noise," *Signal Processing*, vol. 94, pp. 48 – 56, 2014.
- [12] D. Sengupta and S. Kay, "Efficient estimation of parameters for non-gaussian autoregressive processes," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 37, no. 6, pp. 785–794, 1989.
- [13] O. Shalvi and E. Weinstein, "Maximum likelihood and lower bounds in system identification with non-gaussian inputs," *IEEE transactions on information theory*, vol. 40, no. 2, pp. 328–339, 1994.
- [14] J. Lo, "Finite-dimensional sensor orbits and optimal nonlinear filtering," *IEEE Transactions on Information Theory*, vol. 18, no. 5, pp. 583–588, 1972.
- [15] N. I. Achieser, *Theory of Approximation*. Dover Publications, 1992.
- [16] J. K. Tugnait, "Identification of non-minimum phase linear stochastic systems," *Automatica*, vol. 22, no. 4, pp. 457–464, 1986.
- [17] G. B. Giannakis, "On the identifiability of non-gaussian arma models using cumulants," *IEEE Transactions on Automatic Control*, vol. 35, no. 1, pp. 18–26, 1990.
- [18] T. Söderström, *Discrete-time stochastic systems: estimation and control*. Springer Science & Business Media, 2002.
- [19] U. Forssell and L. Ljung, "Identification of unstable systems using output error and box-jenkins model structures," *IEEE Transactions on Automatic Control*, vol. 45, no. 1, pp. 137–141, 2000.
- [20] K. Lii and M. Rosenblatt, "Deconvolution and estimation of transfer function phase and coefficients for non-gaussian linear processes," *Ann. Statist.*, vol. 10, no. 4, pp. 1195–1208, 1982.
- [21] B. Friedlander and B. Porat, "Asymptotically optimal estimation of ma and arma parameters of non-gaussian processes from high-order moments," *IEEE Transactions on Automatic Control*, vol. 35, no. 1, pp. 27–35, 1990.
- [22] B. Porat and B. Friedlander, "Performance analysis of parameter estimation algorithms based on high-order moments," *International Journal of Adaptive Control and Signal Processing*, vol. 3, pp. 191–229, 1989.

- [23] G. B. Giannakis and J. M. Mendel, "Identification of nonminimum phase systems using higher order statistics," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 37, no. 3, pp. 360–377, 1989.
- [24] A. Pandav, D. Mallick, and B. Mohanty, "Effect of limited statistics on higher order cumulants measurement in heavy-ion collision experiments," *Nuclear Physics A*, vol. 991, p. 121608, 2019.
- [25] S. Thil, H. Garnier, and M. Gilson, "Third-order cumulants based methods for continuous-time errors-in-variables model identification," *Automatica*, vol. 44, no. 3, pp. 647 – 658, 2008.
- [26] M. M. Khorasani and E. Weyer, "Non-Asymptotic Confidence Regions for Errors-In-Variables Systems," *IFAC-PapersOnLine*, vol. 51, no. 15, pp. 1020 – 1025, 2018, 18th IFAC Symposium on System Identification.
- [27] R. Carvajal, R. Delgado, J. C. Agüero, and G. C. Goodwin, "An identification method for Errors-in-Variables systems using incomplete data," *IFAC Proceedings Volumes*, vol. 45, no. 16, pp. 1359–1364, 2012.
- [28] A. G. Wills, J. Hendriks, C. Renton, and B. Ninness, "A Bayesian Filtering Algorithm for Gaussian Mixture Models," *ArXiv e-prints*, 2017.
- [29] I. Arasaratnam, S. Haykin, and R. J. Elliott, "Discrete-Time Nonlinear Filtering Algorithms Using Gauss-Hermite quadrature," *Proceedings of the IEEE*, vol. 95, no. 5, pp. 953–977, 2007.
- [30] W.-R. Wu, "Target tracking with glint noise," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 29, pp. 174–185, 01 1993.
- [31] S. Haykin, K. Huber, and Z. Chen, "Bayesian sequential state estimation for MIMO wireless communications," *Proceedings of the IEEE*, vol. 92, pp. 439–454, 2004.
- [32] H. W. Sorenson and D. L. Alspach, "Recursive bayesian estimation using gaussian sums," *Automatica*, vol. 7, no. 4, pp. 465–479, 1971.
- [33] J. Dahlin, A. Wills, and B. Ninness, "Sparse Bayesian ARX models with flexible noise distributions," *IFAC-PapersOnLine*, vol. 51, no. 15, pp. 25 – 30, 2018, 18th IFAC Symposium on System Identification SYSID.
- [34] J. Janouek, P. Gajdo, M. Radecký, and V. Snáel, "Gaussian Mixture Model Cluster Forest," in *IEEE 14th International Conference on Machine Learning and Applications*, 2015, pp. 1019–1023.

- [35] P. A. Torres-Carrasquillo, T. P. Gleason, and D. A. Reynolds, "Dialect identification using Gaussian mixture models," in *ODYSSEY04-The Speaker and Language Recognition Workshop*, 2004.
- [36] R. Orellana, R. Carvajal, and J. C. Agüero, "Maximum Likelihood Infinite Mixture Distribution Estimation Utilizing Finite Gaussian Mixtures," *IFAC-PapersOnLine*, vol. 51, no. 15, pp. 706 – 711, 2018, 18th IFAC Symposium on System Identification.
- [37] J. Kormylo and J. Mendel, "Identifiability of nonminimum phase linear stochastic systems," *IEEE Transactions on Automatic Control*, vol. 28, no. 12, pp. 1081–1090, 1983.
- [38] A. Benveniste, M. Goursat, and G. Ruget, "Robust identification of a nonminimum phase system: Blind adjustment of a linear equalizer in data communications," *IEEE Transactions on Automatic Control*, vol. 25, no. 3, pp. 385–399, 1980.
- [39] D. Donoho, *On minimum entropy deconvolution*. Elsevier, 1981.
- [40] J. Scargle, "Absolute value optimization to estimate phase properties of stochastic time series," *IEEE Transactions on Information Theory*, vol. 23, no. 1, pp. 140–143, 1977.
- [41] K. Åström, P. Hagander, and J. Sternby, "Zeros of sampled systems," *Automatica*, vol. 20, no. 1, pp. 31–38, 1984.
- [42] K. B. Rasmussen, "Maximum likelihood estimation of the parameters of nonminimum phase and noncausal arma models," *IEEE transactions on signal processing*, vol. 42, no. 1, pp. 209–211, 1994.
- [43] U. Forssell and H. Hjalmarsson, "Maximum likelihood estimation of models with unstable dynamics and nonminimum phase noise zeros," *IFAC Proceedings Volumes*, vol. 32, no. 2, pp. 3784–3789, 1999.
- [44] A. P. Dempster, N. M. Laird, and D. Rubin, "Maximum Likelihood from Incomplete Data via the EM Algorithm," *Royal Statistical Society*, vol. 39, no. 1, pp. 1–38, 1977.
- [45] K. Mengersen, C. Robert, and M. Titterton, *Mixtures: Estimation and Applications*. Wiley, 2011.
- [46] S. Gibson and B. Ninness, "Robust maximum-likelihood estimation of multivariable dynamic systems," *Automatica*, vol. 41, no. 10, pp. 1667–1682, 2005.
- [47] S. Gibson, A. Wills, and B. Ninness, "Maximum-likelihood parameter estimation of bilinear systems," *IEEE Transactions on Automatic Control*, vol. 50, no. 10, pp. 1581–1596, 2005.

- [48] R. B. Gopaluni, "A particle filter approach to identification of nonlinear processes under missing observations," *The Canadian Journal of Chemical Engineering*, vol. 86, no. 6, pp. 1081–1092, 2008.
- [49] J. C. Agüero, W. Tang, J. I. Yuz, R. Delgado, and G. C. Goodwin, "Dual time-frequency domain system identification," *Automatica*, vol. 48, no. 12, pp. 3031 – 3041, 2012.
- [50] S. Frühwirth-Schnatter, G. Celeux, and C. Robert, *Handbook of Mixture Analysis*. New York: Chapman and Hall/CRC, 2018.
- [51] R. Carvajal, R. Orellana, D. Katselis, P. Escárate, and J. C. Agüero, "A data augmentation approach for a class of statistical inference problems," *PLOS ONE*, vol. 13, no. 12, pp. 1–24, 12 2018.
- [52] G. U. Yule, "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*, vol. 226, no. 636-646, pp. 267–298, 1927.
- [53] G. B. Giannakis, "Wavelet phase reconstruction using cumulants." in *Digest-International Geoscience and Remote Sensing Symposium (IGARSS)*. IEEE, 1987, pp. 37–41.
- [54] D. R. Brillinger, "The comparison of least squares and third-order periodogram procedures in the estimation of bifrequency," *Journal of Time Series Analysis*, vol. 1, no. 2, pp. 95–102, 1980.
- [55] G. B. Giannakis and J. M. Mendel, "Approximate realization and model reduction of non-minimum phase stochastic systems," in *1986 25th IEEE Conference on Decision and Control*. IEEE, 1986, pp. 1079–1084.
- [56] J. Tugnait, "Identification of linear stochastic systems via second-and fourth-order cumulant matching," *IEEE Transactions on Information Theory*, vol. 33, no. 3, pp. 393–407, 1987.
- [57] J. M. Mendel *et al.*, "Minimum-variance and maximum-likelihood deconvolution for noncausal channel models," *IEEE Transactions on Geoscience and Remote Sensing*, no. 6, pp. 797–808, 1985.
- [58] G. B. Giannakis and J. M. Mendel, "Tomographic wavelet estimation via higher order statistics," in *SEG Technical Program Expanded Abstracts 1986*. Society of Exploration Geophysicists, 1986, pp. 514–516.

- [59] A. Benveniste, M. Goursat, and G. Ruget, "Robust identification of a nonminimum phase system: Blind adjustment of a linear equalizer in data communications," *IEEE Transactions on Automatic Control*, vol. 25, no. 3, pp. 385–399, 1980.
- [60] A. W. Lohmann, G. Weigelt, and B. Wirnitzer, "Speckle masking in astronomy: triple correlation theory and applications," *Applied Optics*, vol. 22, no. 24, pp. 4028–4037, 1983.
- [61] A. K. Jain, "Advances in mathematical models for image processing," *Proceedings of the IEEE*, vol. 69, no. 5, pp. 502–528, 1981.
- [62] M. D. Godfrey, "An exploratory study of the bi-spectrum of economic time series," *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, vol. 14, no. 1, pp. 48–69, 1965.
- [63] M. Lanne and P. Saikkonen, "Modeling expectations with noncausal autoregressions," *Available at SSRN 1210122*, 2008.
- [64] Y. C. Kim and E. J. Powers, "Digital bispectral analysis and its applications to nonlinear wave interactions," *IEEE transactions on plasma science*, vol. 7, no. 2, pp. 120–131, 1979.
- [65] J. D. Scargle, "Studies in astronomical time series analysis. i-modeling random processes in the time domain," *The Astrophysical Journal Supplement Series*, vol. 45, pp. 1–71, 1981.
- [66] J. W. Taylor, "Short-term electricity demand forecasting using double seasonal exponential smoothing," *Journal of the Operational Research Society*, vol. 54, no. 8, pp. 799–805, 2003.
- [67] M. R. Hassan and B. Nath, "Stock market forecasting using hidden markov model: a new approach," in *5th International Conference on Intelligent Systems Design and Applications (ISDA'05)*. IEEE, 2005, pp. 192–196.
- [68] M. Pajovic and J. C. Preisig, "Performance analysis of the least squares based lti channel identification algorithm using random matrix methods," in *2011 49th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*. IEEE, 2011, pp. 516–523.
- [69] B. M. Van Doren and K. G. Horton, "A continental system for forecasting bird migration," *Science*, vol. 361, no. 6407, pp. 1115–1118, 2018.
- [70] D. Balcan, B. Gonçalves, H. Hu, J. J. Ramasco, V. Colizza, and A. Vespignani, "Modeling the spatial spread of infectious diseases: The global epidemic and mobility computational model," *Journal of computational science*, vol. 1, no. 3, pp. 132–145, 2010.

- [71] D. Fanelli and F. Piazza, “Analysis and forecast of covid-19 spreading in china, italy and france,” *Chaos, Solitons & Fractals*, vol. 134, p. 109761, 2020.
- [72] M. E. Salgado, J. I. Yuz, and R. A. Rojas, *Análisis de sistemas lineales*. Pearson Educación, 2005.
- [73] T. B. Schön, A. Wills, and B. Ninness, “System identification of nonlinear state-space models,” *Automatica*, vol. 47, no. 1, pp. 39–49, 2011.
- [74] J. C. Agüero, G. C. Goodwin, and J. I. Yuz, “System identification using quantized data,” in *46th IEEE Conference on Decision and Control*, 2007, pp. 4263–4268.
- [75] B. I. Godoy, G. C. Goodwin, J. C. Agüero, D. Marelli, and T. Wigren, “On identification of FIR systems having quantized output data,” *Automatica*, vol. 47, no. 9, pp. 1905–1915, 2011.
- [76] G. C. Goodwin, S. F. Graebe, M. E. Salgado *et al.*, *Control system design*. Upper Saddle River, NJ: Prentice Hall, 2001.
- [77] T. D. Söderström and P. G. Stoica, *Instrumental variable methods for system identification*. Springer, 1983, vol. 57.
- [78] G. J. McLachlan and T. Krishnan, *The EM algorithm and extensions*. John Wiley & Sons, 2007, vol. 382.
- [79] T. Schön and F. Lindsten, “Learning of dynamical systems-particle filters and markov chain methods,” *Draft available*, 2015.
- [80] R. L. Burden and J. D. Faires, *Numerical Analysis*. Cole, Belmont, CA, 1997.
- [81] M. DeGroot and M. Schervish, *Probability and Statistics*. Addison-Wesley, 2012.
- [82] A. Wald, “Note on the consistency of the maximum likelihood estimate,” *The Annals of Mathematical Statistics*, vol. 20, no. 4, pp. 595–601, 1949.
- [83] H. Cramér, *Mathematical Methods of Statistics*. Asia Publishing House, 1962.
- [84] L. Ljung, “Prediction error estimation methods,” *Circuits, Systems and Signal Processing*, vol. 21, no. 1, pp. 11–21, 2002.
- [85] —, “Prediction error estimation methods,” Linköping University, Tech. Rep., October 2001.
- [86] K. Åström, “Maximum likelihood and prediction error methods,” *Automatica*, vol. 16, no. 5, pp. 551–574, 1980.

- [87] D. Crisan and B. Rozovskii, *The Oxford handbook of nonlinear filtering*. Oxford University Press, 2011.
- [88] V. V. Fedorov, *Theory of optimal experiments*. Elsevier, 2013.
- [89] C. R. Rojas, J. S. Welsh, G. C. Goodwin, and A. Feuer, “Robust optimal experiment design for system identification,” *Automatica*, vol. 43, no. 6, pp. 993–1008, 2007.
- [90] B. I. Godoy, P. E. Valenzuela, C. R. Rojas, J. C. Agüero, and B. Ninness, “A novel input design approach for systems with quantized output data,” in *European Control Conference (ECC)*. IEEE, 2014, pp. 1049–1054.
- [91] J. L. Doob, *Stochastic processes*. New York Wiley, 1953, vol. 101.
- [92] H. Cramer and M. R. Leadbetter, *Stationary and related stochastic processes: Sample function properties and their applications*. Courier Corporation, 2013.
- [93] A. Papoulis and S. U. Pillai, *Probability, random variables, and stochastic processes*. Tata McGraw-Hill Education, 2002.
- [94] T. Kailath, *Linear systems*. Prentice-Hall Englewood Cliffs, NJ, 1980, vol. 156.
- [95] A. K. Tangirala, *Principles of system identification: theory and practice*. Crc Press, 2018.
- [96] P. Eykhoff and K. Åström, *System identification*. Wiley New York, 1974, vol. 14.
- [97] K. Åström and T. Söderström, “Uniqueness of the maximum likelihood estimates of the parameters of an arma model,” *IEEE transactions on automatic control*, vol. 19, no. 6, pp. 769–773, 1974.
- [98] L. Ljung, “Convergence analysis of parametric identification methods,” *IEEE Transactions on Automatic Control*, vol. 23, no. 5, pp. 770–783, 1978.
- [99] —, “Asymptotic behavior of the extended kalman filter as a parameter estimator for linear systems,” *IEEE Transactions on Automatic Control*, vol. 24, no. 1, pp. 36–50, 1979.
- [100] K. Pearson, “Contributions to the mathematical theory of evolution,” *Philosophical Transactions of the Royal Society of London. A*, vol. 185, pp. 71–110, 1894.
- [101] B. D. O. Anderson and J. B. Moore, *Optimal Filtering*. Englewood Cliffs, NJ: Prentice-Hall, 1979.

-
- [102] L. Ljung and T. Söderström, *Theory and practice of recursive identification*. MIT press, 1983.
- [103] P. Gutowski, E. Robinsor, and S. Treitel, “Spectral estimation: Fact or fiction,” *IEEE Transactions on Geoscience Electronics*, vol. 16, no. 2, pp. 80–84, 1978.
- [104] A. Feuer and G. Goodwin, *Sampling in digital signal processing and control*. Springer Science & Business Media, 2012.
- [105] M. M’Saad, R. Ortega, and I. Landau, “Adaptive controllers for discrete-time systems with arbitrary zeros: An overview,” *Automatica*, vol. 21, no. 4, pp. 413–423, 1985.
- [106] R. Orellana, P. Escarate, M. Cure, A. Christen, R. Carvajal, and J. C. Agüero, “A method to deconvolve stellar rotational velocities-iii. the probability distribution function via maximum likelihood utilizing finite distribution mixtures,” *Astronomy & Astrophysics*, vol. 623, p. A138, 2019.
- [107] M. Abramowitz and I. A. Stegun, *Handbook of mathematical functions with formulas, graphs, and mathematical tables*. US Government printing office, 1948, vol. 55.
- [108] J. F. Kenney, *Mathematics of statistics*. D. Van Nostrand, 1939.
- [109] R. A. Fisher, “Moments and product moments of sampling distributions,” *Proceedings of the London Mathematical Society*, vol. 2, no. 1, pp. 199–238, 1930.
- [110] D. Brillinger, “An introduction to polyspectra,” *Ann. Math. Stat.*, vol. 36, no. 5, pp. 1351–1374, 1965.
- [111] D. Brillinger and M. Rosenblatt, “Computation and interpretation of kth order spectra,” *Spectral Analysis of Time Series*, pp. 189–232, 1967.
- [112] —, “Asymptotic theory of estimates of kth order spectra,” *Spectral Analysis of Time Series*, 1967.
- [113] R. A. Davis, K.-S. Lii, and D. N. Politis, *Selected Works of Murray Rosenblatt*. Springer Science & Business Media, 2011.
- [114] K. J. Åström and B. Wittenmark, *Computer-controlled systems: theory and design*. Courier Corporation, 2013.
- [115] G. B. Giannakis, “A kronecker product formulation of the cumulant based realization of stochastic systems,” in *1988 American Control Conference*. IEEE, 1988, pp. 2096–2101.

- [116] B. Friedlander and B. Porat, "The modified yule-walker method of arma spectral estimation," *IEEE Transactions on Aerospace and Electronic Systems*, no. 2, pp. 158–173, 1984.
- [117] J. K. Tugnait, "Spectral estimation for noisy signals observed through a linear system," in *The 22nd IEEE Conference on Decision and Control*. IEEE, 1983, pp. 1331–1336.
- [118] B. Anderson, J. Moore, and R. Hawkes, "Model approximations via prediction error identification," *Automatica*, vol. 14, no. 6, pp. 615–622, 1978.
- [119] K. Åström and T. Söderström, "Uniqueness of the maximum likelihood estimates of the parameters of an arma model," *IEEE Transactions on Automatic Control*, vol. 19, no. 6, pp. 769–773, 1974.
- [120] A. Atkinson, "Likelihood ratios, posterior odds and information criteria," *Journal of Econometrics*, vol. 16, no. 1, pp. 15–20, 1981.
- [121] H. Akaike, "A new look at the statistical model identification," *IEEE Transactions on Automatic Control*, vol. 19, no. 6, pp. 716–723, 1974.
- [122] R. Kashyap, "Inconsistency of the aic rule for estimating the order of autoregressive models," *IEEE Transactions on Automatic Control*, vol. 25, no. 5, pp. 996–998, 1980.
- [123] B. Porat and B. Friedlander, "Optimal estimates of ma and arma parameters of non-gaussian processes from high-order cumulants," in *Fourth Annual ASSP Workshop on Spectrum Estimation and Modeling*. IEEE, 1988, pp. 208–212.
- [124] G. B. Giannakis and J. M. Mendel, "Identification of nonminimum phase systems using higher order statistics," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 37, no. 3, pp. 360–377, 1989.
- [125] A. K. Nandi, *Blind estimation using higher-order statistics*. Springer Science & Business Media, 2013.
- [126] A. Stuart, J. K. Ord, S. F. Arnold, and M. G. Kendall, *Classical inference and the linear model*. Arnold, 2007.
- [127] C. Rose, "k-statistics: Unbiased estimators of cumulants," *Mathematical Statistics with Mathematica*, 2002.
- [128] B. Porat and B. Friedlander, "Performance analysis of parameter estimation algorithms based on high-order moments," *International Journal of adaptive control and signal processing*, vol. 3, no. 3, pp. 191–229, 1989.

- [129] K. Lii, *Proceedings of the IEEE Signal Processing Workshop on Higher-Order Statistics*. IEEE Computer Society, 1997.
- [130] A. Swami, G. B. Giannakis, and G. Zhou, "Bibliography on higher-order statistics," *Signal processing*, vol. 60, no. 1, pp. 65–126, 1997.
- [131] A. Swami, J. Mendel, and C. Nikias, "Hosa - higher order spectral analysis toolbox," *MATLAB Central*, 2003.
- [132] R. J. Kozick and B. M. Sadler, "Maximum-likelihood array processing in non-Gaussian noise with Gaussian mixtures," *IEEE Transactions on Signal Processing*, vol. 48, no. 12, pp. 3520–3535, 2000.
- [133] S. Kalyani and K. Giridhar, "OFDM channel estimation in the presence of NBI and the effect of misspecified NBI model," in *IEEE 8th Workshop on Signal Processing Advances in Wireless Communications*, 2007, pp. 1–5.
- [134] S. Kalyani, "On CRB for Parameter Estimation in Two Component Gaussian Mixtures and the Impact of Misspecification," *IEEE Transactions on Communications*, vol. 60, no. 12, pp. 3734–3744, 2012.
- [135] I. S. Gradshteyn and I. M. Ryzhik, *Table of integrals, series, and products*, 8th ed. Academic press, Elsevier Inc, 2014.
- [136] J. L. Willis, "Acceleration of Generalized Hypergeometric Functions Through Precise Remainder Asymptotics," *Numer. Algorithms*, vol. 59, no. 3, pp. 447–485, 2012.
- [137] J. Guillera and J. Sondow, "Double integrals and infinite products for some classical constants via analytic continuations of Lerch's transcendent," *The Ramanujan Journal*, vol. 16, no. 3, pp. 247–270, 2008.
- [138] P. E. Caines, *Linear Stochastic Systems*. New York, NY, USA: John Wiley & Sons, Inc., 1987.
- [139] I. Bilik and J. Tabrikian, "Target tracking in glint noise environment using nonlinear non-Gaussian Kalman filter," in *IEEE Conference on Radar*, 2006, p. 6.
- [140] A. Bacharoglou, "Approximation of probability distributions by convex mixtures of gaussian measures," *Proceedings of the American Mathematical Society*, vol. 138, no. 7, pp. 2619–2628, 2010.
- [141] D. L. Alspach, "A Gaussian sum approach to the multi-target identification-tracking problem," *Automatica*, vol. 11, no. 3, pp. 285 – 296, 1975.

- [142] D. L. Alspach and H. W. Sorenson, "Nonlinear Bayesian Estimation Using Gaussian Sum Approximations," *IEEE Trans. Automatic Control*, vol. AC-17, no. 4, pp. 439–448, 1972.
- [143] Y. Huang, K. B. Englehart, B. Hudgins, and A. D. C. Chan, "A gaussian mixture model based classification scheme for myoelectric control of powered upper limb prostheses," *IEEE Transactions on Biomedical Engineering*, vol. 52, no. 11, pp. 1801–1811, 2005.
- [144] A. J. Zeevi and R. Meir, "Density estimation through convex combinations of densities: approximation and estimation bounds," *Neural Networks*, vol. 10, no. 1, pp. 99–109, 1997.
- [145] J. Q. Li and A. R. Barron, "Mixture density estimation," in *Advances in neural information processing systems*, 2000, pp. 279–285.
- [146] H. E. Rauch, F. Tung, and C. T. Striebel, "Maximum likelihood estimates of linear dynamic systems," *AIAA journal*, vol. 3, no. 8, pp. 1445–1450, 1965.
- [147] R. M. Neal, "Slice sampling," *Annals of statistics*, pp. 705–741, 2003.
- [148] C. Audet and J. E. Dennis, Jr., "Analysis of Generalized Pattern Searches," *SIAM J. on Optimization*, vol. 13, no. 3, pp. 889–903, 2002.
- [149] W. Peng, "Model selection for Gaussian Mixture Model based on desirability level criterion," *Optik*, vol. 130, pp. 797 – 805, 2017.
- [150] A. Oppenheim and R. Schafer, *Discrete-time signal processing*. Prentice Hall, 1989.
- [151] C. Jin, Y. Zhang, S. Balakrishnan, M. J. Wainwright, and M. I. Jordan, "Local maxima in the likelihood of gaussian mixture models: Structural results and algorithmic consequences," in *Advances in neural information processing systems*, 2016, pp. 4116–4124.
- [152] G. J. McLachlan and D. Peel, *Finite mixture models*. New York: John Wiley & Sons, 2000.
- [153] S. Wright, "Coordinate descent algorithms," *Mathematical Programming*, vol. 151, no. 1, pp. 3–34, Jun 2015.
- [154] S. Boyd, S. P. Boyd, and L. Vandenberghe, *Convex optimization*. Cambridge university press, 2004.
- [155] L. Xu and M. Jordan, "On Convergence properties of the EM Algorithm for Gaussian Mixtures," *Neural Computation*, vol. 8, no. 1, pp. 129–151, 1996.

- [156] I. Naim and D. Gildea, "Convergence of the EM Algorithm for Gaussian Mixtures with Unbalanced Mixing Coefficients," in *Proceedings of the 29th International Conference on International Conference on Machine Learning*, 2012, pp. 1427–1431.
- [157] R. Zhao, Y. Li, and Y. Sun, "Statistical convergence of the EM algorithm on Gaussian mixture models," *Electron. J. Statist.*, vol. 14, no. 1, pp. 632–660, 2020.
- [158] J. Ma, L. Xu, and M. Jordan, "Asymptotic Convergence Rate of the EM Algorithm for Gaussian Mixtures," *Neural Computation*, vol. 12, no. 12, pp. 2881–2907, 2000.
- [159] A. Polanski, M. Marczyk, M. Pietrowska, P. Widlak, and J. Polanska, "Initializing the EM Algorithm for Univariate Gaussian, Multi-Component, Heteroscedastic Mixture Models by Dynamic Programming Partitions," *International Journal of Computational Methods*, vol. 15, no. 03, p. 1850012, 2018.
- [160] N. Hsu and F. J. Breidt, "Exact Maximum Likelihood estimation for non-Gaussian moving averages," *Statistica Sinica*, vol. 19, no. 2, pp. 545–560, 2009.
- [161] J. C. Agüero, J. I. Yuz, G. C. Goodwin, and R. A. Delgado, "On the equivalence of time and frequency domain maximum likelihood estimation," *Automatica*, vol. 46, no. 2, pp. 260–270, 2010.
- [162] T. Schön, *An explanation of the expectation maximization algorithm*. Automatic Control at Linköpings universitet, 2009.
- [163] B. Ninness, "Some system identification challenges and approaches," *IFAC Proceedings Volumes*, vol. 42, no. 10, pp. 1–20, 2009.
- [164] B. Ninness and A. Wills, "An identification toolbox for profiling novel techniques," *IFAC Proceedings Volumes*, vol. 39, no. 1, pp. 919–924, 2006.
- [165] B. Ninness, A. Wills, and A. Mills, "Unit: A freely available system identification toolbox," *Control Engineering Practice*, vol. 21, no. 5, pp. 631–644, 2013.
- [166] E. M. Hendrix, G. Boglárika *et al.*, *Introduction to nonlinear and global optimization*. Springer, 2010, vol. 37.
- [167] M. Ali, C. Storey, and A. Törn, "Application of stochastic global optimization algorithms to practical problems," *Journal of Optimization Theory and Applications*, vol. 95, no. 3, pp. 545–563, 1997.
- [168] A. Törn, M. M. Ali, and S. Viitanen, "Stochastic global optimization: Problem classes and solution techniques," *Journal of Global Optimization*, vol. 14, no. 4, pp. 437–447, 1999.

- [169] S. Mirjalili and A. Lewis, "The whale optimization algorithm," *Advances in engineering software*, vol. 95, pp. 51–67, 2016.
- [170] J. Brownlee, *Clever algorithms: nature-inspired programming recipes*. Jason Brownlee, 2011.
- [171] X. Yao, Y. Liu, and G. Lin, "Evolutionary programming made faster," *IEEE Transactions on Evolutionary computation*, vol. 3, no. 2, pp. 82–102, 1999.
- [172] S. Mirjalili and A. Lewis, "S-shaped versus v-shaped transfer functions for binary particle swarm optimization," *Swarm and Evolutionary Computation*, vol. 9, pp. 1–14, 2013.
- [173] R. Horst, P. M. Pardalos, and N. Van Thoai, *Introduction to global optimization*. Springer Science & Business Media, 2000.
- [174] D. Bertsekas and A. Nedic, *Convex analysis and optimization*. Athena Scientific, 2003.
- [175] Z. Michalewicz and D. B. Fogel, *How to solve it: modern heuristics*. Springer Science & Business Media, 2013.
- [176] F. W. Glover and G. A. Kochenberger, *Handbook of metaheuristics*. Springer Science & Business Media, 2006, vol. 57.
- [177] E.-G. Talbi, *Metaheuristics: from design to implementation*. John Wiley & Sons, 2009, vol. 74.
- [178] C. Blum and A. Roli, "Metaheuristics in combinatorial optimization: Overview and conceptual comparison," *ACM computing surveys (CSUR)*, vol. 35, no. 3, pp. 268–308, 2003.
- [179] V. Granville, M. Krivánek, and J.-P. Rasson, "Simulated annealing: A proof of convergence," *IEEE transactions on pattern analysis and machine intelligence*, vol. 16, no. 6, pp. 652–656, 1994.
- [180] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi, "Optimization by simulated annealing," *science*, vol. 220, no. 4598, pp. 671–680, 1983.
- [181] S. Kirkpatrick, "Optimization by simulated annealing: Quantitative studies," *Journal of statistical physics*, vol. 34, no. 5-6, pp. 975–986, 1984.
- [182] L. Ingber, "Simulated annealing: Practice versus theory," *Mathematical and computer modelling*, vol. 18, no. 11, pp. 29–57, 1993.

-
- [183] R. Azencott, *Simulated annealing: parallelization techniques*. Wiley-Interscience, 1992, vol. 27.
- [184] P. J. van Laarhoven and E. Aarts, *Simulated Annealing: Theory and Applications*. Kluwer Academic Publishers, 1988.
- [185] C. R. P. Valenzuela and H. Hjalmarsson, “A simulated annealing approach to exact experiment design for dynamical systems,” in *Reglermöte 2014, Linköping University*, 2014.
- [186] D. E. Goldberg, K. Deb, J. H. Clark *et al.*, “Genetic algorithms, noise, and the sizing of populations,” *Complex Systems Champaign*, vol. 6, 1992.
- [187] T. Bäck, “Optimal mutation rates in genetic search,” in *Proceedings of the fifth international conference on genetic algorithms*. Citeseer, 1993.
- [188] H. Muhlenbein, “How genetic algorithms really work: I. mutation and hillclimbing,” in *Proc. 2nd Int. Conf. on Parallel Problem Solving from Nature, 1992*. Elsevier, 1992.
- [189] J. H. Holland, “Information processing in adaptive systems,” in *Information Processing in the Nervous System: Proceedings of the International Union of Physiological Sciences*, vol. 3, 1962, pp. 330–339.
- [190] —, “Outline for a logical theory of adaptive systems,” *Journal of the ACM (JACM)*, vol. 9, no. 3, pp. 297–314, 1962.
- [191] —, “Adaptive plans optimal for payoff-only environments(adaptive algorithms and reproductive plans for optimal payoff economics),” in *Second Hawaii Conference on Systems Sciences*, 1969.
- [192] R. B. Hollstien, “Artificial genetic adaptation in computer control systems,” Ph.D. dissertation, The University of Michigan, 1972.
- [193] J. H. Holland *et al.*, *Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control, and artificial intelligence*. MIT press, 1992.
- [194] D. Whitley, “A genetic algorithm tutorial,” *Statistics and computing*, vol. 4, no. 2, pp. 65–85, 1994.
- [195] D. E. Goldenberg, *Genetic algorithms in search, optimization and machine learning*. Addison Wesley, Reading: MA, 1989.

- [196] D. E. Goldberg *et al.*, *The Design of Innovation: Lessons from and for Competent Genetic Algorithms* by David E. Goldberg. Springer Science & Business Media, 2002, vol. 7.
- [197] K. Kristinsson and G. A. Dumont, "System identification and control using genetic algorithms," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 22, no. 5, pp. 1033–1046, 1992.
- [198] R. Eberhart and J. Kennedy, "A new optimizer using particle swarm theory," in *MHS'95. Proceedings of the Sixth International Symposium on Micro Machine and Human Science*. IEEE, 1995, pp. 39–43.
- [199] J. Kennedy, "Small worlds and mega-minds: effects of neighborhood topology on particle swarm performance," in *Proceedings of the 1999 congress on evolutionary computation-CEC99 (Cat. No. 99TH8406)*, vol. 3. IEEE, 1999, pp. 1931–1938.
- [200] J. Kennedy and R. Eberhart, "Particle swarm optimization," in *Proceedings of ICNN'95-International Conference on Neural Networks*, vol. 4. IEEE, 1995, pp. 1942–1948.
- [201] Y. Shi and R. Eberhart, "A modified particle swarm optimizer," in *1998 IEEE international conference on evolutionary computation proceedings. IEEE world congress on computational intelligence (Cat. No. 98TH8360)*. IEEE, 1998, pp. 69–73.
- [202] R. Poli, J. Kennedy, and T. Blackwell, "Particle swarm optimization," *Swarm intelligence*, vol. 1, no. 1, pp. 33–57, 2007.
- [203] R. Poli, "Analysis of the publications on the applications of particle swarm optimisation," *Journal of Artificial Evolution and Applications*, vol. 2008, 2008.
- [204] R. C. Eberhart, Y. Shi, and J. Kennedy, *Swarm intelligence*. Elsevier, 2001.
- [205] A. Alfí and H. Modares, "System identification and control using adaptive particle swarm optimization," *Applied Mathematical Modelling*, vol. 35, no. 3, pp. 1210–1221, 2011.
- [206] B. Luitel and G. K. Venayagamoorthy, "Particle swarm optimization with quantum infusion for system identification," *Engineering Applications of Artificial Intelligence*, vol. 23, no. 5, pp. 635–649, 2010.
- [207] J. A. Nelder and R. Mead, "A simplex method for function minimization," *The computer journal*, vol. 7, no. 4, pp. 308–313, 1965.

-
- [208] W. C. Yu, “Positive basis and a class of direct search techniques,” *Scientia Sinica (in Chinese)*, vol. 9, no. S1, pp. 53–67, 1979.
- [209] V. Torczon, “On the convergence of pattern search algorithms,” *SIAM Journal on optimization*, vol. 7, no. 1, pp. 1–25, 1997.
- [210] R. Hooke and T. A. Jeeves, ““direct search” solution of numerical and statistical problems,” *Journal of the ACM (JACM)*, vol. 8, no. 2, pp. 212–229, 1961.
- [211] C. Audet and J. E. Dennis Jr, “Analysis of generalized pattern searches,” *SIAM Journal on optimization*, vol. 13, no. 3, pp. 889–903, 2002.
- [212] R. M. Lewis, V. Torczon, and M. W. Trosset, “Direct search methods: then and now,” *Journal of computational and Applied Mathematics*, vol. 124, no. 1-2, pp. 191–207, 2000.
- [213] C. Conlan, *Automated Trading with R*. Springer, 2016.
- [214] O. Bozorg-Haddad, M. Solgi, and H. A. Loáiciga, *Meta-heuristic and evolutionary algorithms for engineering optimization*. John Wiley & Sons, 2017.
- [215] W.-m. Li, Z.-h. Jiang, T.-l. Wang, and H.-p. Zhu, “Optimization method based on generalized pattern search algorithm to identify bridge parameters indirectly by a passing vehicle,” *Journal of Sound and Vibration*, vol. 333, no. 2, pp. 364–380, 2014.
- [216] J. Herrera, A. Ibeas, M. de la Sen, E. Rivera, and J. Peláez, “Generalized pattern search methods for control of stable, unstable and integrating systems with unknown delay under step input,” *Mathematics and Computers in Simulation*, vol. 115, pp. 37–48, 2015.
- [217] K. Price, R. M. Storn, and J. A. Lampinen, *Differential evolution: a practical approach to global optimization*. Springer Science & Business Media, 2006.

APÉNDICE

A.1. Probabilidades y procesos aleatorios

A continuación se presentan algunas definiciones y resultados básicos de probabilidad y procesos aleatorios.

Definición A.1 (Esperanza).

Sea x una variable aleatoria y $p(x)$ su función de densidad de probabilidad. Entonces, la esperanza de la VA x , que corresponde a su media se encuentra dada por

$$\mu_x = \mathcal{E}\{x\} \triangleq \int_{-\infty}^{\infty} xp(x)dx \quad (\text{A.1})$$

Definición A.2 (Independencia).

Considere 2 variables aleatorias x_1 y x_2 . Dichas variables aleatorias son independientes, $x_1 \perp\!\!\!\perp x_2$, si

$$p(x_1, x_2) = p(x_1)p(x_2) \quad (\text{A.2})$$

Definición A.3 (Probabilidad condicional).

La densidad de probabilidad condicional de la VA x dado y se define como

$$p(x|y) \triangleq \frac{p(x, y)}{p(y)} \quad (\text{A.3})$$

Definición A.4 (Regla de la cadena).

A partir de (A.3) se obtiene la regla de la cadena (de probabilidad), la cual se encuentra dada por

$$p(x_n, x_{n-1}, \dots, x_1) = p(x_n|x_{n-1}, x_{n-2}, \dots, x_1) \cdot p(x_{n-1}, x_{n-2}, \dots, x_1) \quad (\text{A.4})$$

Note que dicha propiedad permite obtener la densidad conjunta a partir de condicionales. Al repetir dicho proceso con cada término final de (A.4) se obtiene

$$p(x_n, x_{n-1}, \dots, x_1) = \prod_{i=1}^n p(x_i|x_{1:i-1}) \quad (\text{A.5})$$

Definición A.5 (Desigualdad de Cauchy-Schwartz).

Dadas 2 variables aleatoria x e y , se tiene que la desigualdad de Cauchy-Schwartz se encuentra dada por

$$|\mathcal{E}\{xy\}|^2 \leq \mathcal{E}\{|xy|\}^2 \leq \mathcal{E}\{x^2\} \mathcal{E}\{y^2\} \quad (\text{A.6})$$

Teorema A.1 (Bayes).

$$p(x|y) = \frac{p(y|x)p(x)}{p(y)} \quad (\text{A.7})$$

Note que el teorema de Bayes permite intercambiar condicionantes.

Definición A.6 (Desigualdad de Jensen).

Sea y una variable aleatoria y $f(\cdot)$ una función convexa continua, entonces

$$f(\mathcal{E}\{y\}) \leq \mathcal{E}\{f(y)\} \quad (\text{A.8})$$

Dicho resultado también es aplicable a esperanzas condicionales de la siguiente manera

$$f(\mathcal{E}\{y|x\}) \leq \mathcal{E}\{f(y)|x\} \quad (\text{A.9})$$

Las definiciones y teoremas aquí presentados se pueden revisar en [91, Capítulo 1]. Para mayor información acerca de procesos estocásticos y variables aleatorias se sugiere revisar [83, 91–93].

A.2. Algoritmo maximización de la esperanza EM

El algoritmo de **maximización de la esperanza**, más conocido como Expectation-Maximization o EM, es un algoritmo iterativo que permite solucionar el problema de estimación mediante **máxima verosimilitud** en modelos probabilísticos que involucran **variables latentes** Z^1 .

Resulta instructivo analizar el algoritmo EM como un procedimiento sistemático que consisten en separar un problema difícil en dos nuevos problemas estrechamente vinculados, los cuales son (idealmente) más manejables que el problema original. Esta separación del problema corresponde a la clave del algoritmo EM.

¹El término **variable latente** es adoptado a partir de estadística y se refiere a una variable que no es directamente observada. Debido a esto, las variables latentes deben ser inferidas (mediante un modelo matemático) a partir de otras variables, las cuales sí pueden ser observadas, i.e. medidas. A las variables latentes también se les conoce como **variables ocultas** o **variables no-observables** y particularmente dentro de la literatura de EM como **datos faltantes** o **datos incompletos**

De manera más formal, el algoritmo EM es un método iterativo que alterna entre el cálculo de una esperanza condicional y su posterior maximización, de ahí el nombre maximización de la esperanza. A continuación se presenta el desarrollo del algoritmo EM básico para la identificación de sistemas dinámicos, basado en [162].

Para ello, considere en primer lugar el problema de estimación mediante máxima verosimilitud, abordado previamente en la **Sección 2.3**, el cual se encuentra dado por

$$\hat{\theta}_{ML} \triangleq \arg \max_{\theta \in \Theta} \ell_N(\theta), \quad \ell_N(\theta) = \log p_\theta(Y), \quad (\text{A.10})$$

con $Y = y_{1:N}$. La estrategia del algoritmo EM consiste en separar el problema de estimación mediante **máxima verosimilitud** (A.10) en dos problemas vinculados, los cuales son (idealmente) más manejables que el problema original. Esta separación se lleva a cabo al explotar la estructura inherente al modelo probabilístico, lo cual es explicado en detalle a continuación.

En vez de estudiar la formulación original del estimador de máxima verosimilitud (A.10), el cual involucra $\ell_N(\theta) = \log p_\theta(Y)$, el algoritmo EM se basa en el siguiente logaritmo de la función de verosimilitud conjunta

$$\ell_\theta(Z, Y) = \log p_\theta(Z, Y), \quad (\text{A.11})$$

donde Z corresponde a las variables latentes. La idea central del algoritmo EM es considerar la función logaritmo de la verosimilitud conjunta (A.11) de los datos observados Y y las variables latentes Z .

Si bien este paso resulta algo confuso de entender en un principio, las razones por las cuales se realiza se esclarecerán posteriormente. Una forma de recuperar $\ell_\theta(Y)$ a partir de $\ell_\theta(Z, Y)$ es a partir de la definición de probabilidad condicional

$$p_\theta(Z|Y) = \frac{p_\theta(Z, Y)}{p_\theta(Y)}, \quad (\text{A.12})$$

la cual nos permite establecer la siguiente conexión

$$\underbrace{\log p_\theta(Y)}_{\ell_\theta(Y)} = \underbrace{\log p_\theta(Z, Y)}_{\ell_\theta(Z, Y)} - \log p_\theta(Z|Y) \quad (\text{A.13})$$

Sea θ_k el vector de parámetros estimados de θ para la k -ésima iteración del algoritmo. Entonces, el problema de separación previamente mencionado es obtenido al integrar (A.13) respecto de $p_{\theta_k}(Z|Y)$, mediante lo cual se obtiene

$$\log p_\theta(Y) = \int \log p_\theta(Z, Y) p_{\theta_k}(Z|Y) dZ - \int \log p_\theta(Z|Y) p_{\theta_k}(Z|Y) dZ$$

$$= \underbrace{\mathcal{E}_{\theta_k} \{p_{\theta}(Z, Y) | Y\}}_{\triangle \mathcal{Q}(\theta, \theta_k)} - \underbrace{\mathcal{E}_{\theta_k} \{p_{\theta}(Z | Y) | Y\}}_{\triangle \mathcal{V}(\theta, \theta_k)}. \quad (\text{A.14})$$

Note que en la ecuación anterior se utiliza el hecho de que

$$\int \log p_{\theta}(Y) p_{\theta_k}(Z | Y) dZ = \log p_{\theta}(Y). \quad (\text{A.15})$$

Resulta importante notar que se ha supuesto que las variables latentes son continuas. No obstante, no hay nada que impida derivar el algoritmo EM para variables latentes discretas, la única diferencia es que las integrales en (A.14) deben ser reemplazadas por sumatorias.

A continuación se estudia la diferencia en el logaritmo de la función de verosimilitud $\ell_{\theta}(Y)$, al ser evaluada en dos valores diferentes, dados por θ y θ_k , es decir

$$\ell_{\theta}(Y) - \ell_{\theta_k}(Y) = (\mathcal{Q}(\theta, \theta_k) - \mathcal{Q}(\theta_k, \theta_k)) + (\mathcal{V}(\theta_k, \theta_k) - \mathcal{V}(\theta, \theta_k)), \quad (\text{A.16})$$

donde $\mathcal{Q}(\theta, \theta_k)$ y $\mathcal{V}(\theta, \theta_k)$ corresponden a las definiciones presentes en (A.14). Al analizar $(\mathcal{V}(\theta_k, \theta_k) - \mathcal{V}(\theta, \theta_k))$ con mayor detalle y utilizar la definición de $\mathcal{V}(\theta, \theta_k)$ presente en (A.14) se tiene

$$\begin{aligned} \mathcal{V}(\theta_k, \theta_k) - \mathcal{V}(\theta, \theta_k) &= \int \log \left(\frac{p_{\theta_k}(Z | Y)}{p_{\theta}(Z | Y)} \right) p_{\theta_k}(Z | Y) dZ \\ &= \mathcal{E}_{\theta_k} \left\{ -\log \left(\frac{p_{\theta_k}(Z | Y)}{p_{\theta}(Z | Y)} \right) \middle| Y \right\}. \end{aligned} \quad (\text{A.17})$$

Dado que el logaritmo negativo es una **función convexa** (ver **Definición B.1**) es posible utilizar la **desigualdad de Jensen** (ver **Definición A.6**) para establecer lo siguiente

$$\begin{aligned} \mathcal{E}_{\theta_k} \left\{ -\log \left(\frac{p_{\theta_k}(Z | Y)}{p_{\theta}(Z | Y)} \right) \middle| Y \right\} &\geq -\log \mathcal{E}_{\theta_k} \left\{ \frac{p_{\theta_k}(Z | Y)}{p_{\theta}(Z | Y)} \middle| Y \right\} \\ &= -\log \int p_{\theta}(Z | Y) dZ = 0, \end{aligned} \quad (\text{A.18})$$

lo cual prueba que

$$\mathcal{V}(\theta_k, \theta_k) - \mathcal{V}(\theta, \theta_k) \geq 0. \quad (\text{A.19})$$

Entonces, al utilizar este resultado en (A.16) y elegir un nuevo parámetro, tal que $\mathcal{Q}(\theta, \theta_k) \geq \mathcal{Q}(\theta_k, \theta_k)$, se incrementa la función de verosimilitud o, en el peor de los casos, se mantiene, i.e.

$$\mathcal{Q}(\theta, \theta_k) \geq \mathcal{Q}(\theta_k, \theta_k) \Rightarrow \ell_{\theta}(Y) \geq \ell_{\theta_k}(Y). \quad (\text{A.20})$$

Lo cual sugiere que, al obtener $\mathcal{Q}(\theta, \theta_k)$ utilizando la definición presente en (A.14) y posteriormente maximizar dicho resultado respecto de θ se obtiene una nueva estimación para θ dada por θ_{k+1} . Luego, debido al resultado dado por (A.20) se tiene que esta

nueva estimación produce un valor superior o al menos igual de la función de verosimilitud respecto de la estimación previa, dada por θ_{k+1} . Este procedimiento se repite hasta lograr la convergencia del algoritmo como se puede apreciar a continuación en el **Algoritmo A.2.1**.

Algoritmo A.2.1: Pseudo-código para el algoritmo maximización de la esperanza (EM).

Entrada: La condición de detención del algoritmo **StopCondition()** y la estimación inicial para el vector de parámetros θ_{ini}

Salida : El vector de parámetros estimado θ .

```

1  $k \leftarrow 0$  ;
2  $\theta_0 \leftarrow \theta_{ini}$  ;
3 while  $\neg \text{StopCondition}()$  do
    Expectation (E)-step:
4      $\mathcal{Q}(\theta, \theta_k) \leftarrow \mathcal{E}_{\theta_k} \{ \log p_\theta(Z, Y) | Y \} = \int \log p_\theta(Z, Y) p_{\theta_k}(Z|Y) dZ$  ;

    Maximization (M)-step:
5      $\theta_{k+1} = \arg \max_{\theta} \mathcal{Q}(\theta, \theta_k)$  ;
6      $k \leftarrow k + 1$  ;
7 end
8 return  $\theta$  ;
```

Existen numerosas maneras de abordar la convergencia del algoritmo EM (**StopCondition**). Una manera comúnmente utilizada es monitorear el valor del logaritmo de la función de verosimilitud y declarar la convergencia del algoritmo cuando el aumento cae por debajo de cierto umbral arbitrario $\varepsilon_\ell > 0$ (cuyo valor por defecto es generalmente $\varepsilon_\ell = 10^{-6}$). Es decir

$$|\ell_{\theta_{k+1}}(Y) - \ell_{\theta_k}| \leq \varepsilon_\ell. \quad (\text{A.21})$$

Otra manera de verificar la convergencia del algoritmo es controlar el cambio en el valor de los parámetros entre iteraciones e indicar que el algoritmo ha convergido cuando

$$\|\theta_{k+1} - \theta_k\|_2 \leq \varepsilon_P \quad (\text{A.22})$$

para algún $\varepsilon_P > 0$ predefinido.

A.2.1. Ejemplo numérico

A continuación se presenta un ejemplo sencillo de identificación de sistemas en variables de estado mediante el algoritmo EM (ver [162, Sección 3 y 4]). La ventaja de considerar un ejemplo sencillo es que facilita la comprensión del algoritmo. Considere entonces el siguiente sistema escalar descrito en variables de estado

$$x_{t+1} = \theta x_t + v_t \quad (\text{A.23a})$$

$$y_t = 0,5x_t + e_t, \quad (\text{A.23b})$$

con $v_t \sim \mathcal{N}(0; 0,1)$, $e_t \sim \mathcal{N}(0; 0,1)$ y $v_t \perp e_t$, es decir independientes. Adicionalmente se asume que el estado inicial del sistema es conocido $x_1 = 0$. El parámetro verdadero del sistema se encuentra dado por $\theta = 0,9$.

Entonces, el problema de identificación consiste en determinar θ a partir de las observaciones $Y = \{y_1, y_2, \dots, y_N\}$ mediante el algoritmo EM. Para ello se consideran los estados del sistema $Z = \{x_1, x_2, \dots, x_N\}$ como variables latentes, lo cual tiene sentido, puesto que al conocer los estados del sistema se puede recuperar el parámetro θ fácilmente mediante una regresión lineal.

El desarrollo completo del algoritmo para el ejemplo presentado (es decir, la obtención de la expresión para el **paso-E** y el **paso-M**) se puede revisar en [162, Sección 3], por ende es omitido. El caso general de identificación de sistemas en variables de estado se encuentra en [46].

A continuación se presentan los resultados obtenidos a partir de siete estudios de Monte-Carlo realizados, utilizando 1000 realizaciones de observaciones $Y = \{y_1, y_2, \dots, y_N\}$ en cada uno. La diferencia entre los estudios es el número de muestras N utilizados en cada uno. La inicialización del parámetro a estimar se encuentra dado por $\theta_0 = 0,1$ para todas las realizaciones de Monte-Carlo.

N	100	200	500	1000	2000	5000	10000
$\hat{\theta}$	0,8716	0,8852	0,8952	0,8978	0,8988	0,8996	0,8998

Tabla A.1: Estadísticas de la simulación de Monte-Carlo al estimar θ mediante el algoritmo EM para el sistema (A.23).

Los resultados de la simulación de Monte-Carlo se presentan en la Tabla A.1, donde se encuentra el valor medio de la estimación obtenida $\hat{\theta}$ para diferentes longitudes de datos N . Se puede apreciar la consistencia del algoritmo, dado que al aumentar el número de muestras, el parámetro estimado tiende al valor real dado por $\theta = 0,9$.

A continuación, en la Fig. A.1, se presentan gráficos del logaritmo de la función de

verosimilitud y la función auxiliar para diferentes iteraciones k del algoritmo EM, con la finalidad de ilustrar de mejor manera su comportamiento. Se consideran $N = 1000$ datos para la obtención de dichas figuras.

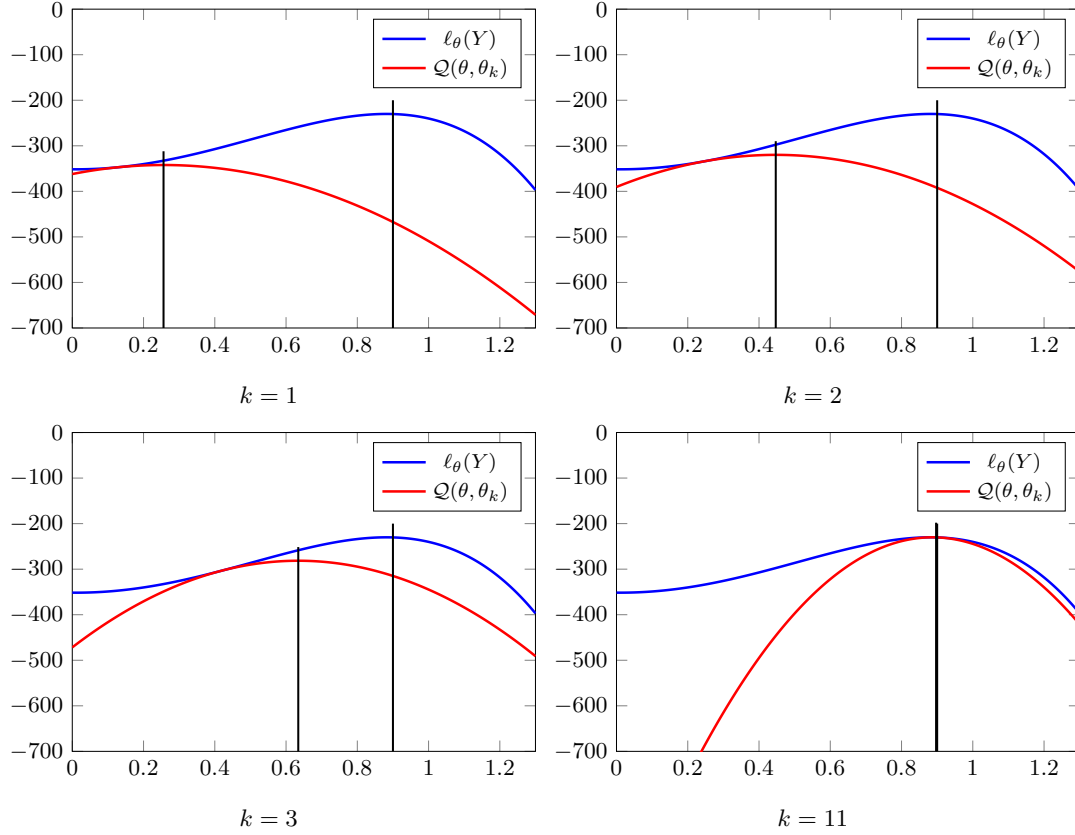


Fig. A.1: Logaritmo de la función de verosimilitud $\ell_\theta(Y)$ y función auxiliar $Q(\theta, \theta_k)$ para diferentes iteraciones k del algoritmo EM para el sistema (A.23)

En la Fig.A.1 se observa el logaritmo de la función de verosimilitud $\ell_\theta(Y)$ para el sistema (A.23). En dichas figuras es posible apreciar el comportamiento de la función auxiliar $Q(\theta, \theta_k)$, obtenida en el *paso-E* del algoritmo. Posteriormente, en el *paso-M* se maximiza dicha función, lo cual se encuentra representado con una línea vertical de color negro. Note la naturaleza creciente (ver (A.20)) de la función auxiliar respecto de las iteraciones del algoritmo.

A.2.2. Lecturas complementarias

- La fuente original del algoritmo se encuentra en [44], donde se proporciona el primer procedimiento sistemático del método y se acuña el nombre maximización de la esperanza.

- Una explicación detallada del algoritmo EM, junto a ejemplos y bibliografía adicional se encuentra en [162].
- En [163] se encuentra una descripción general del algoritmo EM para la identificación del sistemas.
- Una descripción más detallada del algoritmo se presenta en [78], junto a múltiples aplicaciones en diversas áreas.
- La identificación de sistemas en variables de estado mediante el algoritmo EM se puede revisar en [46], mientras que una extensión a sistemas no-lineales se encuentra en [73].
- La utilización del algoritmo EM para la estimación paramétrica de un modelo de mezcla de distribuciones gaussianas se encuentra en [36, 50, 106, 152].
- En [164, 165] se encuentra un toolbox para la identificación de sistemas mediante el algoritmo EM en Matlab y Octave y su documentación asociada.

ALGORITMOS DE OPTIMIZACIÓN GLOBAL

Los problemas de **optimización** son transversales a la mayoría de los campos de estudio, tales como ciencia, ingeniería, matemáticas y física aplicada, tecnologías de la información, entre muchos otros [166–168]. Debido a esto, representa una importante área de investigación al desarrollar nuevos algoritmos de optimización [169, 170] y verificar su comportamiento ante diversos criterios de prueba (e.g. [171, 172]).

B.1. Breve Introducción

La optimización se define (matemáticamente) como *la búsqueda de una combinación de parámetros, o variables de decisión* ($x = x_1, x_2, x_3, \dots, x_n \in \mathbb{R}^n$) *las cuales optimizan (al minimizar o maximizar) una cantidad ordinal* (c) *(usualmente un escalar, el cual se denomina costo) asignada por una función objetivo o función de costo* $f(x)$. *La función objetivo puede encontrarse o no sujeta a un conjunto de restricciones* ($g(x) = \{g_1(x), g_2(x), \dots, g_n(x)\}$), *las cuales corresponden usualmente a limitaciones del área de búsqueda* $x \in X \subseteq \mathbb{R}^n$, *o restricciones dadas por limitaciones de diseño generalmente descritas mediante igualdad* $g_e(x) = 0$ *o de desigualdad* $g_i(x) \leq 0$.

Para mayor información acerca de optimización se sugiere revisar [154, 166, 170, 173, 174]. Cabe destacar los manuales de usuario asociados a los toolbox de optimización y optimización global de MATLAB, puesto que presentan una introducción a distintos algoritmos, los parámetros asociados a éstos, junto con aplicaciones y ejemplos.

Entonces en términos generales, el problema de optimización puede ser escrito como

$$\begin{aligned} &\text{mín } f(x) \text{ sujeto a} \\ &x \in X \\ &g_i(x) \leq 0 \\ &g_e(x) = 0 \end{aligned} \tag{B.1}$$

El objetivo es encontrar el óptimo o mejor plan (al minimizar o maximizar el costo c dado por $f(x)$) para las variables de decisión x . Resulta importante distinguir la diferencia entre un **óptimo local** y un **óptimo global**, lo cual se ilustra en la Fig. B.1

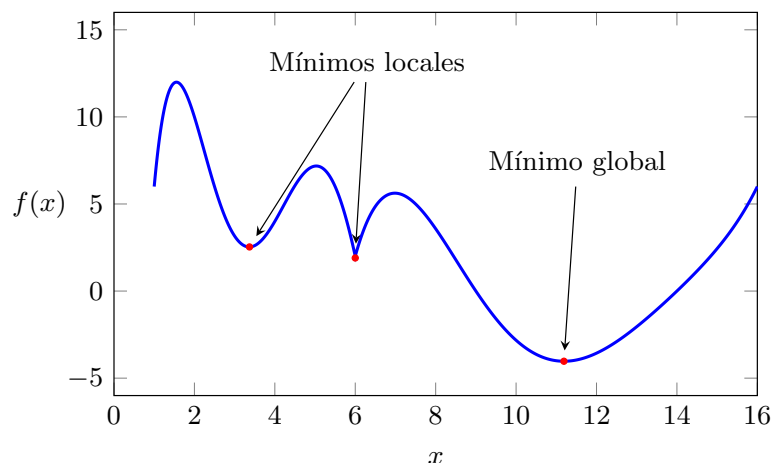


Fig. B.1: Óptimo global y local.

En la Fig. B.1 se puede apreciar un funcional costo $f(x)$ graficado para diversos valores de x dentro de la región de búsqueda o interés X . Denominaremos **óptimo local** al mejor resultado obtenido en una vecindad en torno a él y **óptimo global** al mejor resultado dentro del área de búsqueda X .

Observación B.1. Cabe destacar que el estudio de optimización comprende una gran cantidad de disciplinas especializadas basadas generalmente en una clasificación superpuesta [168, 170, 173]. Por ejemplo, respecto de las variables de decisión, los problemas se pueden clasificar en univariados o multivariados. El tipo de variables de decisión puede ser continuo, discreto o de permutación de variables. Las dependencias entre las variables en el funcional de costo define especialidades como programación lineal, programación cuadrática, programación no-lineal. ∇

Estrategias tradicionales de optimización [154] corresponde a métodos de optimización matemáticos usualmente basados en gradientes (por ejemplo el método de Newton o Descenso de Gradiente que utilizan derivadas para encontrar el mínimo local). Dichos métodos presentan numerosas cualidades atractivas, tales como *efectividad* [166] (propiedad del algoritmo respecto de su habilidad para encontrar el óptimo), *eficiencia* [166] (asociado al esfuerzo requerido para encontrar el óptimo), no obstante se encuentran limitados a una clase particular de problemas denominados **convexos** [154, 174]. A continuación se presenta la definición de **función convexa**:

Definición B.1. La función $f : \mathbb{R}^n \rightarrow \mathbb{R}$ es convexa ssi

$$\forall x_1, x_2 \in \mathbb{R}^n \wedge \forall t \in [0, 1] : \quad f(tx_1 + (1-t)x_2) \leq tf(x_1) + (1-t)f(x_2) \quad (\text{B.2})$$

Dicha definición se puede apreciar de mejor manera en la Fig. B.2

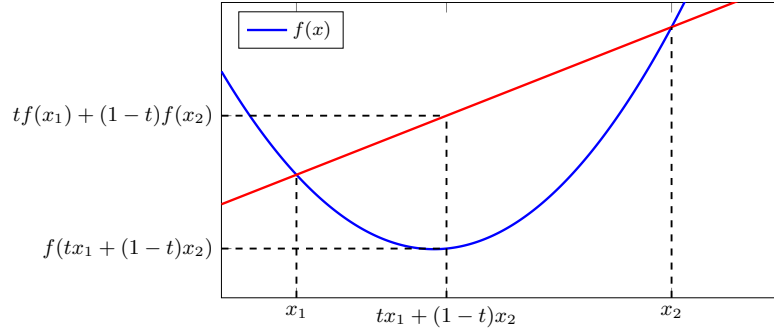


Fig. B.2: Función convexa.

Dichas estrategias poseen limitaciones ante problemas de mayor dificultad, los cuales usualmente poseen múltiples óptimos locales (multi-modales), discontinuidades, no-linealidades, o se encuentran fuertemente restringidos, por mencionar algunas dificultades (en [175] se presenta un análisis detallado de las razones por las cuales los enfoques tradicionales no son adecuados para problemas más complejos).

Por estas y otras razones han surgido principalmente en las últimas décadas estrategias no convencionales denominadas **metaheurísticas**. En este contexto, una **heurística** es un algoritmo que encuentra una solución “suficientemente buena” a un problema, aún cuando no es posible probar si dicha solución es óptima [175].

Los métodos heurísticos poseen generalmente un compromiso en cuanto a precisión, calidad y exactitud en favor del costo computacional.

De manera similar a las heurísticas, las metaheurísticas deben ser consideradas como algoritmos (o estructuras) que pueden ser aplicadas a diferentes problemas de optimización mediante pequeñas modificaciones y ajustes a fin de ser adaptadas a un problema específico [176, 177].

La principal diferencia radica en que las metaheurísticas han sido concebidas tratando de extender las capacidades de las heurísticas al combinar uno o más métodos heurísticos (o procedimientos) mediante una estrategia de alto nivel (por ello el “meta”).

Algunos autores [178] han descrito las propiedades que definen las **metaheurísticas** de la siguiente manera

- Son estrategias que guían el proceso de búsqueda.

- El objetivo es explorar el espacio de búsqueda para encontrar soluciones cuasi-óptimas.
- Las técnicas que constituyen las metaheurísticas varían desde búsquedas locales simples hasta procesos de aprendizaje complejos.
- Los algoritmos son aproximados y usualmente no-determinísticos.
- Generalmente incorporan mecanismos para evitar el quedar atrapados en ciertas áreas del espacio de búsqueda.
- No son específicas para un problema.
- Actualmente las metaheurísticas avanzadas utilizan la experiencia de la búsqueda (incorporada mediante alguna clase de memoria) para guiar la búsqueda.
- Debe poseer un balance entre **exploración** del espacio de búsqueda y **explotación** de óptimos locales encontrados. La explotación se define como la capacidad del algoritmo de optimización para mejorar la mejor solución que ha encontrado hasta ahora buscando en un área pequeña entorno a dicha solución. Mientras que la exploración de un algoritmo, describe la capacidad del algoritmo para salir del óptimo actual y buscar una mejor solución.

Algunos ejemplos de metaheurísticas son la búsqueda local iterada, la búsqueda Tabú, los algoritmos genéticos, la optimización de colonia de hormigas, el recocido simulado, entre muchos otros. A continuación se presentan algunos algoritmos de optimización global heurísticos y un método basado en un patrón de búsqueda. Posteriormente se realiza una evaluación comparativa con la finalidad de encontrar un candidato adecuado para la solución del problema de optimización propuesto en el **Capítulo 4**.

Observación B.2. *Cabe destacar que para maximizar la función objetivo basta con minimizar el negativo de ésta y viceversa.* ∇

B.2. Recocido Simulado (Simulated Annealing)

El recocido simulado (también conocido como Simulated Annealing (SA)) es un algoritmo de optimización global que pertenece a las áreas de optimización estocástica y metaheurísticas.

B.2.1. Inspiración

Este algoritmo se encuentra inspirado en el tratamiento térmico llamado recocido en metalurgia. En dicho proceso el material es calentado hasta cierta temperatura para

posteriormente enfriarlo lentamente en condiciones controladas con la finalidad de permitir la formación de estructuras cristalinas e incrementar el tamaño de estas junto con reducir posibles defectos en el material. El efecto obtenido en el material corresponde a mejoras en cuanto a dureza y durabilidad del material. El calor incrementa la energía de los átomos, permitiendo que se muevan libremente, mientras que el esquema de enfriado paulatino permite la formación de nuevas configuraciones de baja energía (estructuras cristalinas).

B.2.2. Metáfora

Cada candidato a solución en el espacio de búsqueda representa un diferente estado de energía interna del sistema. El aumento de temperatura resulta en una relajación del criterio de aceptación de los posibles candidatos a solución. A medida que el sistema se enfría, el criterio de aceptación de los candidatos se riguriza con el fin de centrarse en encontrar mejores candidatos. Una vez que el sistema se ha enfriado, el candidato a solución representa una muestra en o cercana al óptimo global.

B.2.3. Estrategia y procedimiento

El objetivo es localizar la configuración que posea el menor costo dentro del espacio de búsqueda. La estrategia del algoritmo para llevar a cabo esta tarea consiste en re-muestrear probabilísticamente el espacio de búsqueda en busca de un nuevo candidato. La aceptación de dicho candidato por sobre la muestra actualmente almacenada es administrada por una función probabilística que se vuelve mas exigente o rigurosa a medida que transcurre el tiempo.

Este criterio de decisión probabilístico se basa usualmente en el algoritmo **Metropolis-Hastings** para simular muestras de un sistema termodinámico.

Un pseudo-código generalizado del algoritmo SA para minimizar un funcional de costo se puede apreciar en el **Algoritmo B.2.1**.

Algoritmo B.2.1: Pseudo-código para el algoritmo Recocido Simulado (SA).

Entrada: El tamaño del problema P_{size} , el número máximo de iteraciones N_{iter} , la función de costo $\text{Cost}(\cdot)$, la temperatura máxima $temp_{max}$, el límite superior U (opcional) y el límite inferior l (opcional).

Salida : Mejor solución S_{best} y su costo asociado C_{best} .

```

1  $S_{current} \leftarrow \text{CreateInitialSolution}(P_{size}, U, l)$  ;
2  $C_{current} \leftarrow \text{Cost}(S_{current})$  ;
3  $S_{best} \leftarrow S_{current}$  ;
4  $C_{best} \leftarrow \text{Cost}(S_{best})$  ;
5 for  $i = 1$  to  $N_{iter}$  do
6    $S_i \leftarrow \text{CreateNeighborSolution}(S_{current}, U, l)$  ;
7    $temp_{current} \leftarrow \text{CalculateTemperature}(i, temp_{max})$  ;
8    $C_i \leftarrow \text{Cost}(S_i)$  ;
9   if  $C_i \leq C_{current}$  then
10     $S_{current} \leftarrow S_i$  ;
11     $C_{current} \leftarrow C_i$  ;
12    if  $C_i \leq C_{best}$  then
13       $S_{best} \leftarrow S_i$  ;
14       $C_{best} \leftarrow C_i$  ;
15    end
16   else if  $\text{Exp}(\frac{C_{current} - C_i}{temp_{curr}}) > \text{Rand}()$  then
17      $S_{current} \leftarrow S_i$  ;
18      $C_{current} \leftarrow C_i$  ;
19   end
20 end
21 return  $S_{best}, C_{best}$  ;
```

B.2.4. Algunas consideraciones

- El algoritmo de recocido simulado fue diseñado originalmente para problemas de optimización combinatoria, no obstante se encuentra adaptado para optimización de funciones continuas.
- En [179] se demuestra que para un periodo de enfriamiento suficientemente extenso, el algoritmo converge de manera casi segura hacia el óptimo global. No obstante el número de muestras necesarias para la convergencia puede ser incluso mayor que el número total de muestras presentes en el espacio de búsqueda, por ende dicho

resultado no posee utilidad práctica real.

- El desempeño del algoritmo para un problema particular se puede mejorar mediante un esquema adecuado de selección del nuevo candidato (a través de la función que crea la nueva solución en la vecindad de la actual) que sea menos propenso a generar candidatos con costos más altos.
- Reiniciar el proceso de enfriamiento utilizando la mejor solución encontrada hasta el momento puede conducir a mejores resultados en algunos casos.
- El método de aceptación mas común (no obstante existen varios en la literatura) es aceptar siempre soluciones mejores y aceptar soluciones peores con una probabilidad dada por $P_r\{\text{accept}\} \leftarrow \exp\left(\frac{C-C'}{T}\right)$, donde T representa la temperatura actual, C el costo (o energía) de la solución actual y C' el costo del candidato.
- El tamaño de la vecindad utilizado para la generación del nuevo candidato puede cambiar en el tiempo o encontrarse influenciado por la temperatura, comenzando inicialmente con una vecindad amplia y estrechándose a medida que se ejecuta el algoritmo.
- En muchos casos se utiliza un método heurístico específico para el problema para proveer una solución inicial.

B.2.5. Lecturas complementarias

- La fuente original del algoritmo se encuentra en [180, 181], mientras que la demostración de convergencia se puede revisar en [179].
- Un análisis detallado del método expuesto y mejoras tales como recocido simulado adaptativo, templado simulado y métodos híbridos se encuentra en [182].
- En [183, 184] es posible revisar con mayor detalle la teoría, procedimiento y aplicaciones del método.
- En [185]¹ se encuentra una aplicación del método de recocido simulado en identificación de sistemas, particularmente en el diseño óptimo de entrada.

B.3. Algoritmo genético (Genetic Algorithm)

El algoritmo genético (más conocido por su nombre en inglés, Genetic Algorithm (GA)) corresponde a un estrategia adaptativa y una técnica de optimización global. El

¹Disponible en https://www.control.isy.liu.se/reglermote2014/modules/request.php?module=oc_program&action=summary.php&id=33.

algoritmo genético pertenece al área de computación evolutiva y se encuentra emparejado a una extensa lista de técnicas.

B.3.1. Inspiración

El algoritmo genético se encuentra inspirado, tal como lo implica su nombre, en la genética (incluyendo la herencia y frecuencia de genes) y la teoría evolutiva de la población. Además, incluye la estructura de la genética mendeliana ² (como cromosomas, genes y alelos) y sus mecanismos (recombinación y mutación).

B.3.2. Metáfora

Individuos pertenecientes a una población contribuyen con su material genético (conocido como genotipo) de manera proporcional a la idoneidad de su genoma expresado (denominado fenotipo) a su entorno en forma de descendencia. La siguiente generación de individuos se crea a través de un proceso de apareamiento, el cual implica recombinación de dos genomas individuales de la población junto con la introducción de errores de copia aleatoria (llamada mutación). Este proceso iterativo puede resultar en un mejor ajuste entre los fenotipos de los individuos de una población y su entorno.

B.3.3. Estrategia y procedimiento

El objetivo del algoritmo genético es maximizar los resultados de las soluciones candidatas o individuos pertenecientes a una población respecto de una función de costo. En este contexto cada individuo representa una solución candidata al problema, cuya información se encuentra codificada en su genoma. La estrategia para el Algoritmo Genético es emplear repetidamente los mecanismos genéticos de recombinación y mutación en la población de candidatos a solución. La función de costo es aplicada a la representación decodificada de cada candidato gobernando las contribuciones probabilísticas de supervivencia hacia la siguiente generación, lo cual es denominado mecanismo de selección.

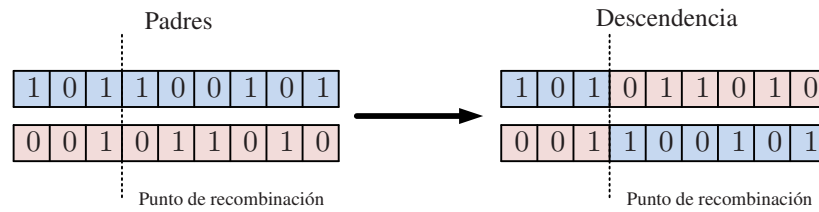
Existen numerosas codificaciones empleadas en algoritmos genéticos para representar las soluciones candidatas como individuos. Las más comunes son codificación binaria y codificación real. Dichos individuos son sometidos a los procesos de recombinación, mutación y selección las cuales describiremos brevemente a continuación.

²debido a Gregor Johann Mendel y sus estudios sobre genética.

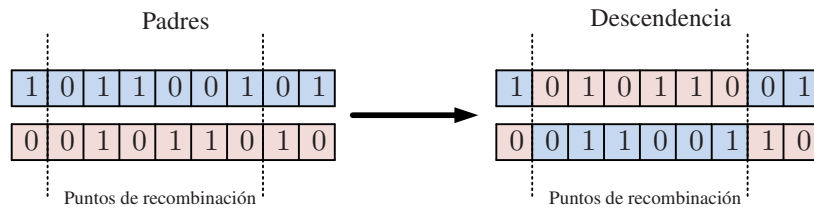
Recombinación (crossover)

La recombinación es un operador genético utilizado para generar variación genética en los individuos de una generación a la siguiente. Es análogo a la recombinación de la reproducción sexual biológica. Existen muchas técnicas de recombinación dependiendo de la codificación utilizada. A continuación revisaremos algunos ejemplos de recombinación binarios (ver Fig. B.3)

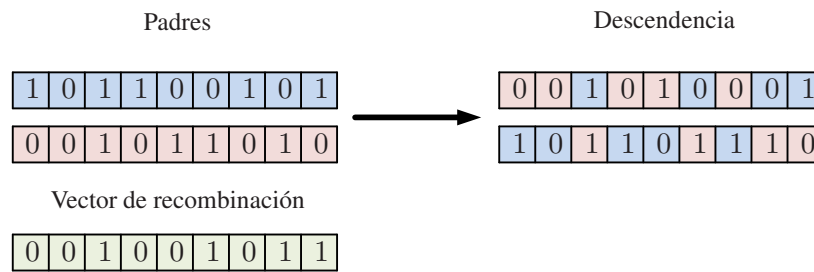
- **Recombinación en un punto (One point crossover):** Se selecciona un punto en el vector del primer parental. Todos los datos ubicados a continuación de dicho punto son intercambiados entre ambos vectores parentales originando de esta manera la descendencia (ver Fig. B.3a).
- **Recombinación en dos puntos (Two points crossover):** De manera análoga a la recombinación en un punto se seleccionan dos puntos en el primer vector parental. Los datos ubicados entre ambos puntos se intercambian entre ambos vectores parentales originando la descendencia (ver Fig. B.3b).
- **Recombinación uniforme (Uniform crossover):** Se crea un vector de recombinación utilizando una distribución de Bernoulli con probabilidad $P_{crossover}$ para cada elemento del vector. Dicho vector indica la procedencia de cada elemento del vector de la descendencia (1 corresponde al primer vector parental y 0 al segundo) (ver Fig. B.3c).



(a) Recombinación en un punto.



(b) Recombinación en dos puntos.



(c) Recombinación uniforme.

Fig. B.3: Ejemplos de recombinación binaria.

Mutación (Mutation)

La mutación es un operador genético usado para mantener la diversidad genética de una población. Es análogo a la mutación biológica. El propósito de la mutación es proveer un mecanismo para escapar de los óptimos locales, así como desplazar a los individuos hacia zonas del espacio de búsqueda que no pueden ser alcanzadas por medio de otros operadores genéticos. El operador de mutación binario tradicionalmente utilizado consiste en la modificación de un bit aleatorio del genoma del individuo seleccionado. Esta modificación se efectúa con una probabilidad preestablecida dada por $P_{mutation}$. En la Fig. B.4 es posible apreciar un ejemplo de mutación binaria.

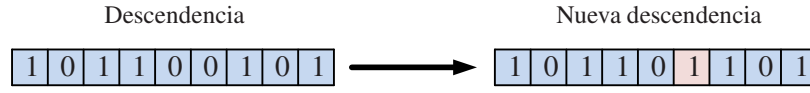


Fig. B.4: Mutación binaria.

Selección (Selection)

La selección es el proceso mediante el cual se eligen individuos a partir de una población, con el objeto de ser reproducidos mediante el operador de recombinación. Este operador es análogo a la selección natural. Si bien existen varios mecanismos de selección, a continuación presentaremos algunos

- **Selección de ruleta (Roulette-wheel Selection):** también conocida como selección proporcional a la función de desempeño. Sea N el número de individuos y f_i el desempeño del i -ésimo individuo (dado por el funcional de costo). La probabilidad asociada a su selección se encuentra dada por

$$p_i = \frac{f_i}{\sum_{j=1}^N f_j} \quad (\text{B.3})$$

Esta selección permite que los mejores individuos sean elegidos con mayor probabilidad, pero al mismo tiempo permite a los peores individuos ser elegidos, lo cual puede ayudar a mantener la diversidad de la población.

- **Selección por truncamiento (Selection by Truncation):** Los individuos son ordenados según su desempeño. Posteriormente se reproduce solo una proporción arbitraria de ellos (predefinida), eliminando los individuos de peor desempeño, lo cual destruye la diversidad y por ende no es muy utilizada en la práctica.
- **Selección por torneo (Tournament Selection):** Efectuada mediante un torneo (comparación basada en el funcional de costo) entre un pequeño subconjunto de tamaño arbitrario de individuos elegidos al azar desde la población.

En el **Algoritmo B.3.1** se presenta un pseudo-código generalizado para el algoritmo genético

Algoritmo B.3.1: Pseudo-código para el algoritmo genético (GA)

Entrada: El tamaño de la población Pop_{size} , el tamaño del problema $Prob_{size}$, la probabilidad de recombinación $P_{crossover}$, la probabilidad de mutación $P_{mutation}$.

Salida : Mejor solución S_{best} y su costo asociado C_{best} .

```

1 Population ← InitializePopulation ( $Pop_{size}, Prob_{size}$ ) ;
2 EvaluatePopulation (Population);
3 ( $S_{best}, C_{best}$ ) ← GetBestSolution (Population);
4 while  $\neg$  StopCondition() do
5     Parents ← SelectParents (Population,  $Pop_{size}$ ) ;
6     Children ←  $\emptyset$  ;
7     foreach  $Parent_1, Parent_2 \in$  Parents do
8          $Child_1, Child_2 \leftarrow$  Crossover ( $Parent_1, Parent_2, P_{crossover}$ ) ;
9         Children ← Mutate ( $Child_1, P_{mutation}$ ) ;
10        Children ← Mutate ( $Child_2, P_{mutation}$ ) ;
11    end
12    EvaluatePopulation( $Children$ ) ;
13    ( $S_{best}, C_{best}$ ) ← GetBestSolution (Children);
14    Population ← Replace (Population, Children)
15 end
16 return  $S_{best}, C_{best}$  ;

```

B.3.4. Algunas consideraciones

- Resulta deseable adoptar representaciones específicas para el problema y operadores genéticos personalizados, incorporando tanta información previa sobre el problema como sea posible.
- El tamaño de la población Pop_{size} debe ser lo suficientemente grande como para proporcionar una cobertura suficiente del dominio del problema y garantizar una adecuada mezcla de los individuos más adecuados [186].
- El algoritmo genético se ha configurado tradicionalmente utilizando una alta probabilidad de recombinación (entre el 95[%]-99[%] de la población seleccionada) y una baja probabilidad de mutación (usualmente cercano al 5[%]) [187, 188].
- En algunas versiones del algoritmo se utiliza elitismo, lo cual se refiere a un número arbitrario (seleccionado por el usuario) de individuos inmunes a los procesos de

selección, recombinación (crossover) y mutación. Este proceso garantiza la supervivencia de las mejores soluciones entre generaciones.

- El proceso de selección de soluciones candidatas proporcional a su aptitud (dada por el funcional de costo) para contribuir a la siguiente generación no debe ser demasiado ambiciosa (debido a que puede evitar la generación de soluciones candidatas más adecuadas) ni demasiado aleatoria.

B.3.5. Lecturas complementarias

- La fuente original del algoritmo genético se encuentra en [189–191].
- En [192] se presentan algunas aplicaciones de control adaptativo y optimización de funciones.
- Algunos resultados, aplicaciones y marco teórico junto con un análisis de las capacidades del algoritmo se presenta en [193].
- Un tutorial detallado sobre los aspectos prácticos y teóricos del algoritmo genético se encuentra en [194].
- Un análisis profundo sobre el método expuesto, aplicaciones y algoritmos desarrollados a partir de éste se puede revisar en [195, 196].
- En [197] se encuentra una aplicación del algoritmo genético en identificación de sistemas y control, particularmente en la estimación de sistemas ARMAX y control adaptativo.

B.4. Optimización mediante enjambre de partículas (Particle Swarm Optimization)

La optimización mediante enjambre de partículas (también llamado Particle Swarm Optimization PSO) es un algoritmo de optimización perteneciente a las áreas de inteligencia computacional, particularmente inteligencias de enjambre y colectivas.

B.4.1. Inspiración

La optimización mediante PSO se inspira en el comportamiento social de alimentación de algunos animales, tales como el comportamiento de bandada en aves y el comportamiento de cardumen en los peces.

B.4.2. Metáfora

Las partículas pertenecientes al enjambre se desplazan a través de un entorno siguiendo a los miembros más calificados (dado por el funcional de costo) del enjambre, dirigiéndose hacia las áreas históricamente mejores del entorno.

B.4.3. Estrategia y procedimiento

El objetivo del algoritmo es concentrar las partículas (que representan soluciones candidatas) en torno al óptimo global del espacio de búsqueda. Para ello distribuye inicialmente las partículas de manera aleatoria en el espacio de búsqueda y les asigna una velocidad inicial pequeña aleatoriamente. Luego el algoritmo se ejecuta de manera iterativa, actualizando la posición de cada partícula en función de su velocidad, el mejor lugar conocido por ella y la mejor posición conocida por una partícula del enjambre. Las partículas son analizadas en cada iteración mediante el funcional de costo. Con el paso del tiempo, a través de una combinación de *exploración* y *explotación* de buenas posiciones conocidas del espacio de búsqueda, las partículas se agrupan o convergen a un óptimo.

El algoritmo PSO está compuesto por una colección de partículas que se mueven en el espacio de búsqueda influenciadas por su mejor ubicación pasada y la mejor ubicación pasada de todo el enjambre. En cada iteración se actualiza la velocidad de una partícula mediante

$$v_i(t+1) = v_i(t) + \left(c_1 \times \text{Rand}() \times (p_i^{best} - p_i(t)) \right) + \left(c_2 \times \text{Rand}() \times (p_{gbest} - p_i(t)) \right) \quad (\text{B.4})$$

donde $v_i(t+1)$ corresponde a la nueva velocidad para la i -ésima partícula, c_1 y c_2 representan los coeficientes de peso para la mejor posición individual y global respectivamente. $p_i(t)$ representa la posición de la i -ésima partícula para el tiempo t , mientras que p_i^{best} corresponde a su mejor posición conocida. La mejor posición conocida por el enjambre se encuentra dada por p_{gbest} . La función $\text{Rand}()$ genera muestras de una variable aleatoria (RV) distribuida de manera uniforme en el intervalo $[0, 1]$, es decir $\sim \mathcal{U}[0, 1]$.

La posición de cada partícula es actualizada mediante

$$p_i(t+1) = p_i(t) + v_i(t) \quad (\text{B.5})$$

Un pseudo-código generalizado para el algoritmo de optimización mediante enjambre

de partículas se presenta a continuación en **Algoritmo B.4.1**.

Algoritmo B.4.1: Pseudo-código para el algoritmo de optimización mediante enjambre de partículas (PSO).

Entrada: El tamaño de la población Pop_{size} , el tamaño del problema $Prob_{size}$ y los coeficientes de peso c_1 y c_2 (mejor posición individual, mejor posición global).

Salida : Mejor solución S_{best} y su costo asociado C_{best} .

```

1 Population  $\leftarrow \emptyset$  ;
2  $P_{g\_best} \leftarrow \emptyset$  ;
3 for  $i = 1$  to  $Pop_{size}$  do
4    $P_{velocity} \leftarrow \text{RandomVelocity}()$  ;
5    $P_{position} \leftarrow \text{RandomPosition}( Prob_{size} )$  ;
6    $P_{cost} \leftarrow \text{Cost}( P_{position} )$  ;
7    $P_{p\_best} \leftarrow P_{position}$  ;
8   if  $P_{cost} \leq \text{Cost}(P_{g\_best})$  then
9      $P_{g\_best} \leftarrow P_{p\_best}$  ;
10  end
11 end
12 while  $\neg \text{StopCondition}()$  do
13   foreach  $P \in \text{Population}$  do
14      $P_{velocity} \leftarrow \text{UpdateVelocity}( P_{velocity}, P_{g\_best}, P_{p\_best}, c_1, c_2 )$  ;
15      $P_{position} \leftarrow \text{UpdatePosition}( P_{position}, P_{velocity} )$  ;
16      $P_{cost} \leftarrow \text{Cost}( P_{position} )$  ;
17     if  $P_{cost} \leq \text{Cost}(P_{p\_best})$  then
18        $P_{p\_best} \leftarrow P_{position}$  ;
19       if  $P_{cost} \leq \text{Cost}(P_{g\_best})$  then
20          $P_{g\_best} \leftarrow P_{position}$  ;
21       end
22     end
23   end
24 end
25  $S_{best} \leftarrow P_{g\_best}$  ;
26  $C_{best} \leftarrow \text{Cost}( P_{g\_best} )$  ;
27 return  $S_{best}, C_{best}$  ;
```

B.4.4. Algunas consideraciones

- La velocidad máxima a la cual una partícula se puede mover (cambio máximo de posición por iteración) debiese ser limitada. Un criterio usual es restringirla a un porcentaje del tamaño del dominio (espacio de búsqueda).
- Los coeficientes de peso c_1 y c_2 asociados a la mejor posición personal y global respectivamente se encuentran generalmente entre 0 a 4, usualmente 2.
- En general se utiliza un número de partículas entre 20-50.
- Existen variantes al Algoritmo B.4.1 en las cuales se introduce un término adicional en (B.4) que representa el momentum o inercia de la partícula, limitando de esta manera el cambio en la velocidad. Otras variantes consideran vecindades de partículas con buenos resultados en (B.4).
- Dado que las partículas pudiesen eventualmente salir del espacio de búsqueda del problema, generalmente se utiliza un coeficiente de penalización o son reflejadas hacia el interior del dominio del problema.

B.4.5. Lecturas complementarias

- La fuente original del algoritmo PSO se puede revisar en [198–200].
- Algunas modificaciones relevantes al método original, como la introducción de la inercia, se encuentra en [201].
- Una perspectiva moderna y diversas extensiones del algoritmo PSO se puede revisar en [202].
- Un análisis centrado en las áreas de aplicación y referencias a estas se presenta en [203].
- En [204] es posible revisar con mayor detalle la teoría, procedimientos y otras técnicas desarrolladas a partir del algoritmo PSO.
- En [205] se encuentra una aplicación de la optimización mediante enjambre de partículas en identificación de sistemas y control, particularmente en la identificación paramétrica de un sistema no-lineal y la optimización de un controlador PID. En [206] el algoritmo es utilizado para identificar los parámetros de un sistema IIR.

B.5. Búsqueda de patrones (Pattern Search)

El algoritmo de búsqueda de patrones (más conocido como Pattern Search (PS)), también denominado búsqueda libre de derivadas o búsqueda directa, es un método de optimización que no requiere información acerca del gradiente de la función objetivo. Este algoritmo pertenece a las áreas de optimización metaheurísticas categorizado como método de búsqueda directa.

A diferencia de los métodos de optimización tradicionales, que utilizan información sobre el gradiente o derivadas para buscar un punto óptimo, el algoritmo PS evalúa un conjunto de puntos alrededor del punto actual, buscando uno donde el valor de la función objetivo sea mejor que el valor en el punto actual.

Dichos puntos ubicados en la vecindad del punto actual son conocidos como rejilla (**mesh**). Dicha rejilla se forma agregando el punto actual a un múltiplo escalar de un conjunto de vectores llamado patrón (**pattern**). Si el algoritmo de búsqueda de patrones encuentra un punto en la rejilla que mejora el costo de la función objetivo en el punto actual, el nuevo punto se convierte en el punto actual en el siguiente paso del algoritmo y el tamaño de la rejilla aumenta. En caso contrario, el punto actual se mantiene y el tamaño de la rejilla disminuye.

El método de búsqueda de patrones es utilizado usualmente para resolver problemas en los cuales la función objetivo es no-diferenciable, o discontinua.

B.5.1. Estrategia y procedimiento

El objetivo del algoritmo es localizar el punto que posea el mejor (mínimo o máximo) costo dentro del espacio de búsqueda. Para llevar a cabo esta tarea, el algoritmo evalúa el funcional de costo en el punto actual, dado por $x(i) \in \mathbb{R}^n$. Cabe destacar que la posición inicial, es decir $x(1) \in \mathbb{R}^n$, es generado de manera aleatoria. Luego genera el conjunto de vectores, denominado patrón como

$$p_j(i) = \{p_k\}_{k=1:n} \in \mathbb{R}^n; j \in [1, n]$$

$$p_k = \begin{cases} 1, & \text{si } k = j \\ 0, & \text{en otro caso.} \end{cases} \quad (\text{B.6})$$

donde n corresponde al tamaño del problema (número de dimensiones de las variables de decisión). Note que se generan n vectores de dimensión n , los cuales al ser apilados forman la matriz identidad $I(n \times n)$.

A partir del vector de patrones se genera la rejilla $M(i)$ en torno al punto actual $x(i)$ de la siguiente manera

$$M(i) = \{m_l(i) : m_l(i) = x(i) \pm \Delta(i)p_j(i), \forall j \in [1, n]\} \quad (\text{B.7})$$

donde $\Delta(i)$ corresponde al tamaño actual de la rejilla. Note que la cantidad de elementos m_l pertenecientes a la rejilla $M(i)$ esta dado por $l = 2n$.

Una vez generados los puntos de la rejilla son evaluados mediante el funcional de costo. El elemento de la rejilla que posea el mejor costo es contrastado con el costo del punto actual. Si dicho punto posee un mejor costo respecto del punto actual, entonces en la siguiente iteración tomara su lugar, es decir $x(i+1) = m_l(i)$ tal que $m_l(i) = \min f(m_l(i)) \forall l \in [1, 2n]$. Junto con el desplazamiento del punto actual, se aumenta el tamaño de la rejilla $\Delta(i+1) = c_{ef}\Delta(i)$, donde $c_{ef} > 1 \in \mathbb{R}^+$ corresponde al coeficiente de expansión de la rejilla.

En caso contrario, es decir si el mejor punto (en términos del funcional de costo) perteneciente a la rejilla posee un peor desempeño que el punto actual, en la siguiente iteración se mantiene el punto actual, es decir $x(i+1) = x(i)$. En este caso se contrae el tamaño de la rejilla, es decir $\Delta(i+1) = c_{cf}\Delta(i)$, donde $c_{cf} < 1 \in \mathbb{R}^+$ representa el coeficiente de contracción de la rejilla.

Lo anteriormente descrito se puede apreciar con mayor claridad en la Fig. B.5. El punto central y los puntos de la grilla se encuentran representados por un triangulo negro y círculos grises respectivamente. El costo asociado a ellos se encuentra al lado derecho de cada posible solución. Note que en la primera iteración (Fig. B.5a) la grilla se desplaza hacia la abajo y se expande (Fig. B.5b). Luego se desplaza nuevamente hacia abajo y expande (Fig. B.5c). En la última iteración se puede apreciar la contracción de la grilla, puesto que el punto central posee el menor costo (Fig. B.5d)

A continuación en el **Algoritmo B.5.1** se puede apreciar un pseudo-código generalizado del algoritmo de optimización mediante búsqueda de patrones.

B.5.2. Algunas consideraciones

- El método de búsqueda de patrones PS se encuentra estrechamente relacionado con numerosos métodos, por ejemplo la búsqueda de la sección dorada (Golden-section search) y el método de Nelder-Mead [207].
- Los coeficientes c_{ef} y c_{cf} , que representan los factores de expansión y contracción respectivamente, se encuentran generalmente dados por $c_{ef} = 2$ y $c_{cf} = 0,5$.
- En [208] se demuestra la convergencia del algoritmo. Posteriormente en [209] se demuestra que el algoritmo converge para una cierta clase de funciones. Fuera de dichas clases el algoritmo es una heurística que puede proporcionar soluciones aproximadas útiles en varios problemas, no obstante puede fallar en algunos.
- Un criterio usualmente utilizado es detener el algoritmo cuando el tamaño de la rejilla es menor que cierto valor predefinido, es decir $\Delta(i) \leq \varepsilon$, $\varepsilon \ll 1 \in \mathbb{R}$.

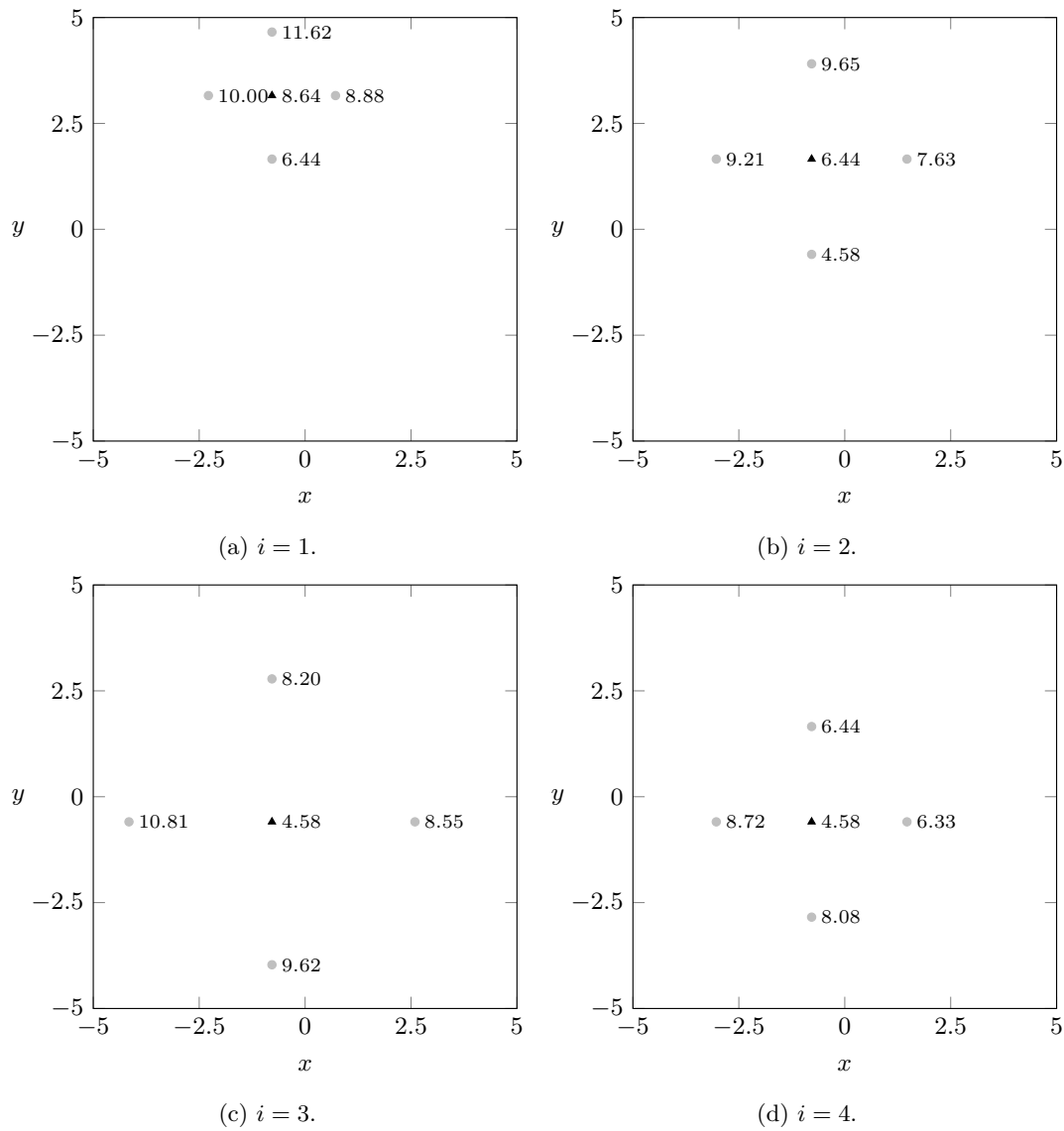


Fig. B.5: Comportamiento del algoritmo de optimización mediante búsqueda de patrones (PS) para diferentes iteraciones.

- El algoritmo propuesto presenta un patrón de búsqueda fijo, definido por el conjunto de vectores dado por (B.6). No obstante existen variaciones en las cuales se rota o cambian los vectores de búsqueda.

B.5.3. Lecturas complementarias

- La fuente original del algoritmo de búsqueda de patrones (PS) se puede revisar en [210], mientras que la demostración de convergencia se puede revisar en [208, 209].

Algoritmo B.5.1: Pseudo-código para el algoritmo de optimización mediante búsqueda de patrones (PS).

Entrada: El tamaño del problema P_{size} , la función de costo $\mathbf{Cost}$, el coeficiente de expansión c_{ef} , el coeficiente de contracción c_{cf} y el tamaño inicial de la rejilla Δ .

Salida : Mejor solución S_{best} y su costo asociado C_{best} .

```

1  $S_{current} \leftarrow \mathbf{CreateInitialSolution}(P_{size})$  ;
2  $C_{current} \leftarrow \mathbf{Cost}( S_{current} )$  ;
3 while  $\neg \mathbf{StopCondition}()$  do
4    $\mathbf{Mesh} \leftarrow \mathbf{CreateMesh}( S_{current}, \Delta )$ ;
5   foreach  $m \in \mathbf{Mesh}$  do
6     if  $\mathbf{Cost}(m) < C_{current}$  then
7        $S_{current} \leftarrow m$  ;
8        $C_{current} \leftarrow \mathbf{Cost}( m )$  ;
9        $\Delta \leftarrow \mathbf{ExpandMesh}( \Delta, c_{ef} )$  ;
10    else
11       $\Delta \leftarrow \mathbf{RefineMesh}( \Delta, c_{cf} )$  ;
12    end
13  end
14 end
15  $S_{best} \leftarrow S_{current}$  ;
16  $C_{best} \leftarrow C_{current}$  ;
17 return  $S_{best}, C_{best}$  ;
```

- En [211] es posible revisar un análisis del método y algoritmos desarrollados a partir de éste.
- Una perspectiva actualizada y algunas extensiones del método PS se pueden revisar en [212].
- En [213,214] es posible revisar con mayor detalle la teoría, procedimientos y marco teórico del algoritmo.
- En [215] se presenta una aplicación del algoritmo de búsqueda de patrones en identificación de sistemas, particularmente en la identificación paramétrica del modelo de un puente. En [216] el algoritmo es utilizado para identificar el retardo de un sistema y ajustar los parámetros de un controlador.

B.6. Comparación entre algoritmos

En el campo de la optimización es común el someter los algoritmos a evaluaciones comparativas de su desempeño. Esto se realiza mediante un conjunto de funciones matemáticas, que presentan diferentes características, cuyo óptimo global es conocido (ver e.g. [217, Capítulo 3] y [171, Tabla 1]). Existen diversos criterios de desempeño, por ejemplo el tiempo de ejecución del algoritmo, la media y desviación estándar obtenida a partir de una simulación de Monte-Carlo, el número de evaluaciones del funcional de costo, el número de iteraciones, entre muchos otros.

B.6.1. Funciones de prueba

A continuación revisaremos algunas funciones clásicas de prueba junto a sus definiciones, soluciones óptimas y áreas usuales de búsqueda. Las definiciones de las funciones de prueba descritas pueden diferir levemente de las definiciones originales. Note que las pruebas se encuentran formuladas como problemas de minimización, no obstante se pueden utilizar como problemas de maximización al invertir el signo de la función (ver **Observación B.2**).

Función de Matyas

La función de Matyas constituye una prueba bastante sencilla para los algoritmos de optimización, puesto que se trata de una función continua, convexa y unimodal. Dicha función posee la siguiente definición

$$f(x, y) = 0,26(x^2 + y^2) - 0,48xy \quad (\text{B.8})$$

El área de prueba se encuentra usualmente restringida al espacio $-10 \leq x, y \leq 10$. El mínimo global se encuentra dado por $f(x, y) = 0$ con $(x, y) = (0, 0)$.

Función de Easom

La función de Easom corresponde a una función unimodal cuyo mínimo global se encuentra en un área pequeña en comparación al espacio de búsqueda. Dicha función se encuentra definida por

$$f(x, y) = -\cos(x) \cos(y) e^{-(x-\pi)^2 - (y-\pi)^2} \quad (\text{B.9})$$

El área de búsqueda se acota usualmente a $-100 \leq x, y \leq 100$. El mínimo global se encuentra dado por $f(x, y) = -1$ y se obtiene cuando $(x, y) = (\pi, \pi)$.

Función de Rosenbrock

La función de Rosenbrock (también conocida como la función banana) constituye un problema de optimización clásico. El óptimo global se encuentra en un valle largo y estrecho con forma parabólica. Si bien encontrar el valle resulta trivial, la convergencia hacia el óptimo global es bastante difícil por lo cual es utilizado con frecuencia para medir el desempeño de los algoritmos de optimización. La definición de la función de Rosenbrock se encuentra dada por

$$f(x_1, \dots, x_n) = \sum_{i=1}^{n-1} \left[100(x_{i+1} - x_i^2)^2 + (1 - x_i)^2 \right] \quad (\text{B.10})$$

El área de búsqueda se encuentra usualmente restringida al hipercubo dado por $-2 \leq x_i \leq 2$ con $i = 1, \dots, n$.

El mínimo global corresponde a $f(x_1, \dots, x_n) = 0$ y se obtiene cuando $x_i = 1$, $i = 1, \dots, n$.

Función de Rastrigin

La función de Rastrigin se encuentra basada en la función de esfera ($f(x_1, \dots, x_n) = \sum_{i=1}^{n-1} [x_i^2]$) con la adición de un coseno, el cual produce varios mínimos locales. Si bien dicha función es altamente multimodal, la ubicación de los mínimos se encuentra uniformemente distribuida. La función de Rastrigin se encuentra definida por

$$f(x_1, \dots, x_n) = 10n + \sum_{i=1}^n [x_i^2 - 10 \cos(2\pi x_i)] \quad (\text{B.11})$$

El área de prueba se encuentra acotada por $-5,12 \leq x_i \leq 5,12$, $i = 1, \dots, n$. El mínimo global se encuentra dado por $f(x_1, \dots, x_n) = 0$ y se obtiene cuando $x_i = 0$, $i = 1, \dots, n$.

Función de Ackley

La función de Ackley es una función multimodal ampliamente usada para probar algoritmos de optimización. Se encuentra determinada por

$$f(x_1, \dots, x_n) = -a \cdot e^{-b \sqrt{\frac{1}{n} \sum_{i=1}^n [x_i^2]}} - e^{\frac{1}{n} \sum_{i=1}^n [\cos(cx_i)]} + a + e \quad (\text{B.12})$$

Los parámetros habitualmente utilizados son $a = 20$, $b = 0,2$, $c = 2\pi$. El área de búsqueda se acota usualmente a $-5 \leq x_i \leq 5$, $i = 1, \dots, n$.

El mínimo global se encuentra dado por $f(x_1, \dots, x_n)$ y se obtiene cuando $x_i = 0$, $i = 1, \dots, n$.

Funciones engañosas (Deceptive)

Se denomina función engañosa a una clase de problemas en los cuales la región total correspondiente a los mínimos locales es considerablemente mayor a la región del óptimo global. A partir de lo anterior es posible inferir que se trata de funciones multimodales. La forma general de las funciones engañosas se puede apreciar a continuación

$$f(x_1, \dots, x_n) = - \left[\frac{1}{n} \sum_{i=1}^n [g_i(x_i)] \right]^\beta \quad (\text{B.13})$$

Donde β corresponde a un factor no lineal fijo.

En la literatura se han definido al menos 3 tipos de problemas engañosos dependiendo de la forma de $g_i(x_i)$. Un problema engañoso del Tipo III corresponde a una función en la cual el óptimo global se encuentra en $x_i = \alpha_i$, $i = 1, \dots, n$, donde α_i corresponde a un número aleatorio tal que $0 < \alpha_i < 1$. Para ello se propone la siguiente función auxiliar

$$g_i(x_i) = \begin{cases} -\frac{x_i}{\alpha_i} + \frac{4}{5} & \text{si } 0 \leq x_i \leq \frac{4}{5}\alpha_i \\ \frac{5x_i}{\alpha_i} - 4 & \text{si } \frac{4}{5}\alpha_i < x_i \leq \alpha_i \\ \frac{5(x_i - \alpha_i)}{\alpha_i - 1} + 1 & \text{si } \alpha_i < x_i \leq \frac{1+4\alpha_i}{5} \\ \frac{x_i - 1}{1 - \alpha_i} + \frac{4}{5} & \text{si } \frac{1+4\alpha_i}{5} < x_i \leq 1 \end{cases} \quad (\text{B.14})$$

Los otros tipos de funciones engañosas corresponden a casos especiales en los cuales $\alpha_i = 1$, $i = 1, \dots, n$ (Tipo I) o $\alpha_i \in \{0, 1\}$ aleatoriamente (Tipo II) para cada dimensión $i = 1, \dots, n$.

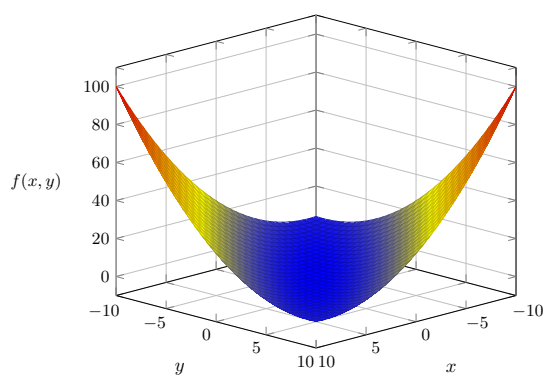
Observación B.3. La función dada por (B.14) debe ser ajustada de manera adecuada para problemas del Tipo I y II. ∇

Para los 3 tipos de funciones engañosas, el tamaño de la región en la cual se encuentran los óptimos locales es $5^n - 1$ veces mayor que la región en la cual se encuentra el óptimo global.

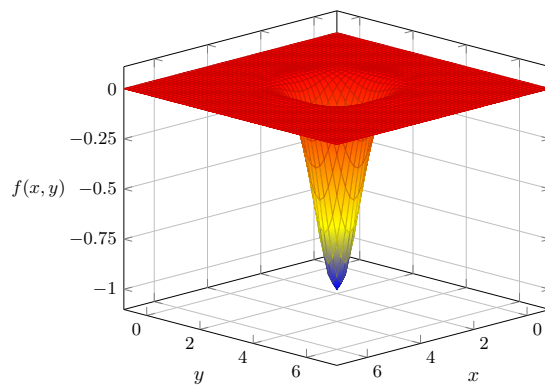
El número de óptimos locales corresponde a $2^n - 1$ para los problemas de Tipo I y II y $3^n - 1$ para los problemas de Tipo III.

El área de prueba se encuentra acotada por $0 \leq x_i \leq 1$, $i = 1, \dots, n$.

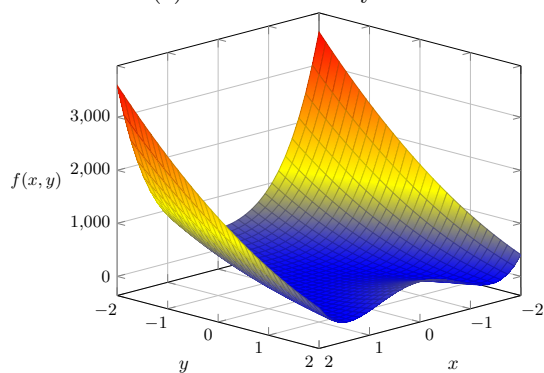
El mínimo global corresponde a $f(x_1, \dots, x_n) = -1$ y se obtiene cuando $x_i = \alpha_i$, $i = 1, \dots, n$.



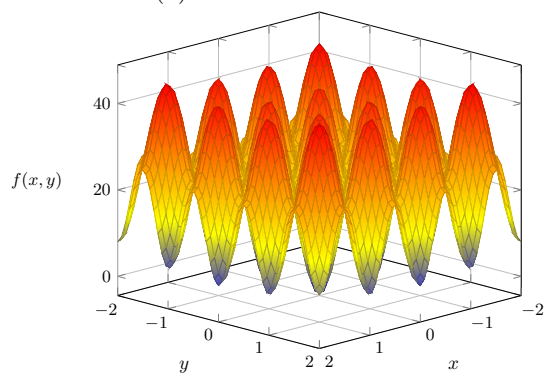
(a) Función de Matyas.



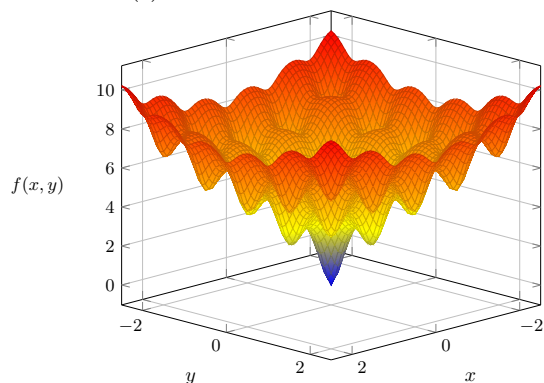
(b) Función de Easom.



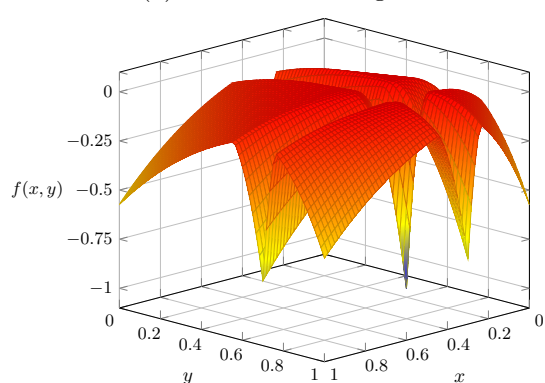
(c) Función de Rosenbrock.



(d) Función de Rastrigin.



(e) Función de Ackleys.

(f) Función engañosa del Tipo III con $\alpha_1 = 0,3$, $\alpha_2 = 0,7$ y $\beta = 2,5$.Fig. B.6: Funciones de prueba utilizadas en $\mathbb{R}^2$.

Las funciones de Matyas, Easom y Rosenbrock son unimodales (poseen solo un óptimo global sin óptimos locales), por ende son apropiadas para evaluar las características de **explotación** de los métodos propuestos. Mientras que las funciones de Rastrigin, Ackley y engañosa poseen múltiples óptimos locales, por lo cual resultan adecuadas para evaluar

las capacidades de **exploración** de los algoritmos y sus capacidades de escape de los óptimos locales.

B.6.2. Pruebas realizadas en $\mathbb{R}^2$

Utilizando las funciones definidas previamente (ver **Subsección B.6.1**) se prueban los algoritmos SA, GA, PSO y PS.

La configuración de la simulación es la siguiente:

- (1) La temperatura máxima utilizada en el algoritmo SA corresponde a $temp_{max} = 1000$.
- (2) El tamaño del enjambre utilizado en el algoritmo PSO esta dado por $Pop_{size} = 250$, mientras que los coeficientes de peso empleados corresponden a $c_1 = 2,5$ y $c_2 = 1,5$.
- (3) El número de individuos utilizados el algoritmo GA esta dado por $Pop_{size} = 100$. Las técnicas de selección de ruleta y recombinación uniforme son empleadas, con una probabilidad de recombinación dada por $P_{crossover} = 0,6$. La probabilidad de mutación es $P_{mutation} = 0,015$ y el número de individuos élite es $n_{elite} = 4$. Se utiliza una codificación binaria con 5 decimales.
- (4) Los coeficientes de expansión y contracción usados en el algoritmo PS se encuentran dados por $c_{ef} = 2$ y $c_{cf} = 0,5$ respectivamente.
- (5) El número de simulaciones de Monte Carlo (MC) es de 50 para cada función de prueba y algoritmo.
- (6) Se considera un número máximo de iteraciones para el algoritmo SA de $N_{iter} = 10000$ y $N_{iter} = 1000$ para los demás.
- (7) Las soluciones iniciales se generan de forma aleatoria dentro del espacio de búsqueda.

El promedio de los resultados obtenidos luego de 50 ejecuciones independientes del algoritmo se pueden apreciar en las Tablas B.1 a B.3 a continuación

Algorithm	Matyas	Easom	Rosenbrock	Rastrigin	Ackleys	Deceptive
SA	6.24	6.25	6.57	6.55	6.86	8.93
PSO	4.34	4.32	4.61	4.64	4.96	6.16
PS	0.02	0.02	0.02	0.02	0.03	0.05
GA	1.52	1.53	1.61	1.63	1.65	1.76

Tabla B.1: Tiempos de ejecución promedio $t_e [s]$.

Algorithm	Matyas	Easom	Rosenbrock	Rastrigin	Ackleys	Deceptive
SA	1.51E-10	1.75E-10	6.37E-12	0	8.88E-16	2.22E-8
PSO	0	0	3.90E-21	0	8.88E-16	3.87E-3
PS	7.89E-1	8.09E-1	6.25E-3	0	2.45E-15	3.66E-2
GA	1.66E-3	1.70E-3	6.81E-4	2.99E-8	6.24E-3	5.15E-4

Tabla B.2: Error promedio μ_e .

Algorithm	Matyas	Easom	Rosenbrock	Rastrigin	Ackleys	Deceptive
SA	1.93E-10	2.60E-10	3.81E-11	0	0	1.33E-8
PSO	0	0	1.96E-20	0	0	5.97E-3
PS	4.09E-1	3.94E-1	1.31E-2	0	2.43E-15	1.29E-2
GA	1.23E-3	1.18E-3	1.19E-3	1.27E-7	4.52E-3	9.39E-4

Tabla B.3: Desviación estándar del error σ_e .

En la Tabla B.1 se observan los tiempos de ejecución promedio de los diferentes algoritmos. Note que el algoritmo de búsqueda de patrones PS posee el menor tiempo de ejecución entre los algoritmos presentados. Esto se debe en parte a que es el algoritmo que realiza una menor cantidad de evaluaciones del funcional (en total).

En los resultados expuestos en la Tabla B.2 y la Tabla B.3 se puede apreciar que la totalidad de los algoritmos descritos en éste capítulo son capaces de solucionar los problemas de optimización asociados a las funciones de prueba con resultados aceptables. Además es posible apreciar que poseen un buen balance entre *exploración* y *explotación* junto con una adecuada capacidad de evadir óptimos locales.

No obstante, el problema de optimización en el cual nos encontramos interesados (dado por (4.21) en el **Capítulo 4**) es multidimensional y multimodal, por ende resulta adecuado evaluar el comportamiento de los algoritmos en dichas condiciones.

B.6.3. Pruebas realizadas en $\mathbb{R}^n$; $n \in [2, 20]$

Utilizando la misma configuración para la simulación (ver **Subsección B.6.2**) se realizan pruebas sobre las funciones de Rosenbrock, Rastrigin y Ackley modificando la dimensión del espacio de búsqueda entre 2 y 20 (note que las definiciones dadas por (B.10), (B.11) y (B.12) son multidimensionales). Los resultados obtenidos se pueden

apreciar a continuación en las Figs. B.7 a B.9

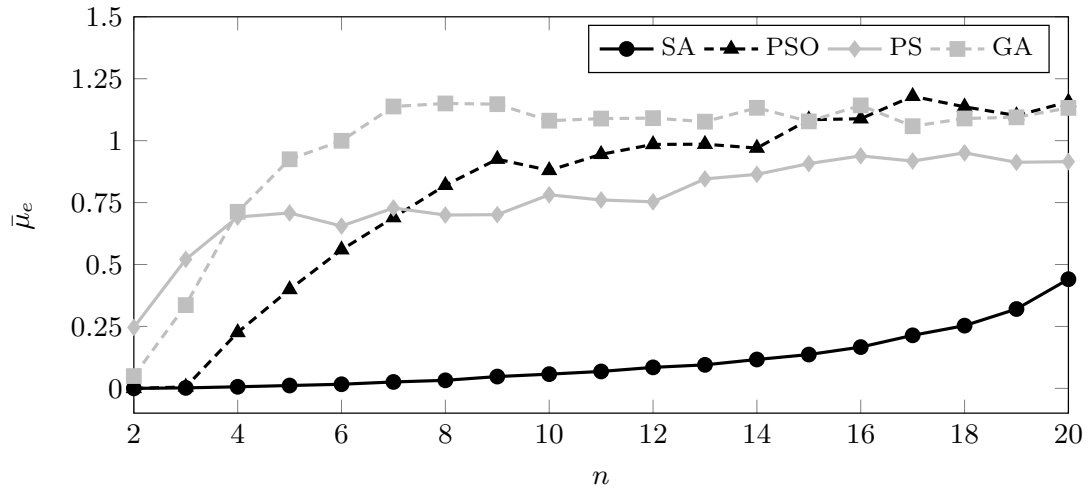


Fig. B.7: Error medio $\bar{\mu}_e$ de las soluciones óptimas obtenidas utilizando diferentes algoritmos sobre la función de Rosenbrock en $\mathbb{R}^n$, $n \in [2, 20]$.

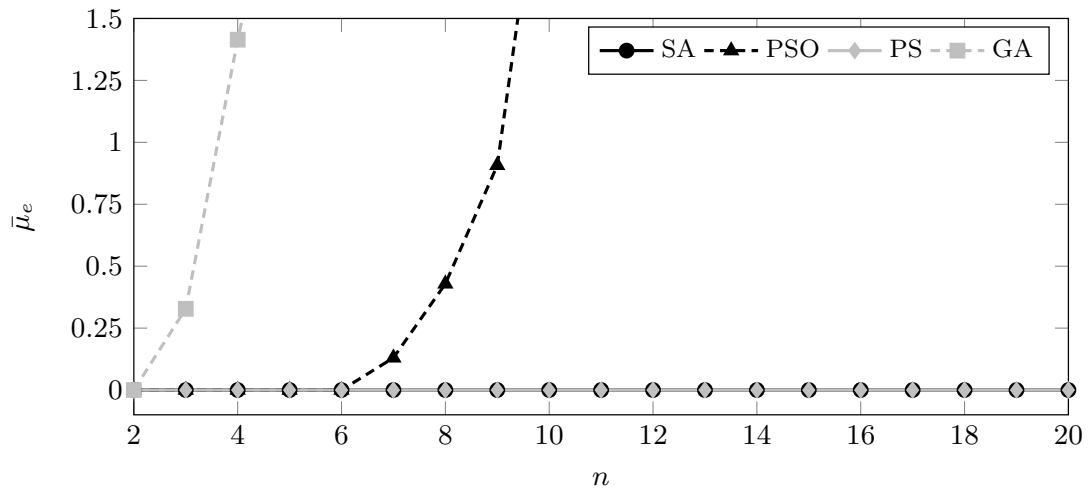


Fig. B.8: Error medio $\bar{\mu}_e$ de las soluciones óptimas obtenidas utilizando diferentes algoritmos sobre la función de Rastrigin en $\mathbb{R}^n$, $n \in [2, 20]$.

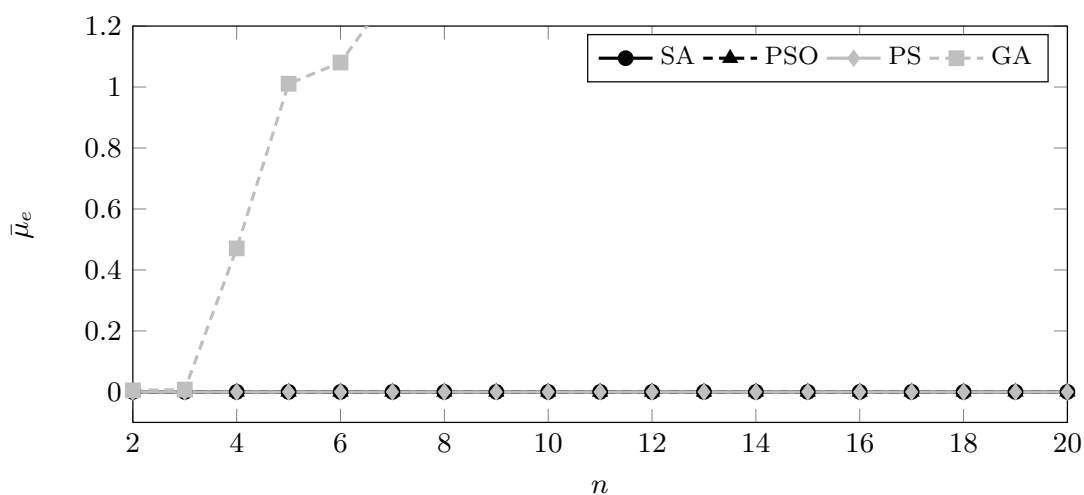


Fig. B.9: Error medio $\bar{\mu}_e$ de las soluciones óptimas obtenidas utilizando diferentes algoritmos sobre la función de Ackley en $\mathbb{R}^n$, $n \in [2, 20]$.

Algorithm	Rosenbrock	Rastrigin	Ackleys
SA	6.736	7.416	8.272
PSO	3.890	4.672	4.988
PS	0.030	0.038	0.041
GA	1.947	2.510	2.465

Tabla B.4: Tiempos de ejecución promedio t_e [s].

A partir de los resultados obtenidos en la simulación de Monte-Carlo (ver Fig. B.7, Fig. B.8 y Fig. B.9) se puede apreciar que los algoritmos de recocido simulado SA y búsqueda de patrones PS poseen el mejor desempeño, en cuanto a error medio, dentro de los métodos propuestos.

Adicionalmente en la Tabla B.4 es posible apreciar los tiempos medios de ejecución de cada algoritmo. Sin lugar a dudas, el algoritmo de búsqueda de patrones PS, presenta el mejor desempeño en cuanto a tiempo de ejecución frente a los demás algoritmos. Por otra parte, el algoritmo que presenta el mayor tiempo de ejecución medio es el de recocido simulado (SA).

Debido a que el método de optimización basado en búsqueda de patrones (PS) es capaz de encontrar soluciones cercanas al óptimo y posee el menor tiempo de ejecución de los algoritmos expuestos, resulta un candidato razonable para solucionar el problema de

optimización en el cual nos encontramos interesados. Dicho método es capaz de solucionar problemas multimodales y multidimensionales en el menor tiempo, comparado con los demás algoritmos presentados. Además es el método más simple de ajustar al problema, dado que solo posee 2 parámetros de ajuste (el coeficiente de expansión y el de contracción de la rejilla).

B.7. Códigos utilizados

A continuación se encuentran hipervínculos hacia algunos códigos de MATLAB utilizados en el desarrollo de la tesis.

B.7.1. Recocido simulado (Simulated Annealing)

El código para el algoritmo de recocido simulado (Simulated Annealing) se encuentra disponible en <https://www.dropbox.com/s/c76f2kjh562iot2/SA.m?dl=0>

B.7.2. Algoritmo genético (Genetic Algorithm)

Codificación Binaria (Binary coded)

El código para el algoritmo genético con codificación binaria (Binary coded Genetic Algorithm) se encuentra disponible en <https://www.dropbox.com/s/y7hrq0smf8dnm3q/BGA.m?dl=0>

Codificación Real (Real coded)

El código para el algoritmo genético con codificación real (Real coded Genetic Algorithm) se encuentra disponible en <https://www.dropbox.com/s/omp61ur1thmqasy/RGA.m?dl=0>

B.7.3. Búsqueda de patrones (Pattern Search)

El código para el algoritmo de búsqueda de patrones (Pattern Search) se encuentra disponible en <https://www.dropbox.com/s/lvwcv7rl6cveeqt/GPS.m?dl=0>

B.7.4. Optimización mediante enjambre de partículas (Particle Swarm Optimization)

El código para el algoritmo de optimización mediante enjambre de partículas (Particle Swarm Optimization) se encuentra disponible en <https://www.dropbox.com/s/8r5rl7ypunefajb/PSO.m?dl=0>

