

*Monitoreo de Perfiles No Lineales Multivariados usando
un enfoque de Datos Funcionales*

JUAN CARLOS ESPINOSA MORENO
ESTADÍSTICO



UNIVERSIDAD NACIONAL DE COLOMBIA
FACULTAD DE CIENCIAS
DEPARTAMENTO DE ESTADÍSTICA
BOGOTÁ, D.C.
2021

*Monitoreo de Perfiles No Lineales Multivariados usando
un enfoque de Datos Funcionales*

JUAN CARLOS ESPINOSA MORENO
ESTADÍSTICO

DISERTACIÓN PRESENTADA PARA OPTAR AL TÍTULO DE
MAGISTER EN CIENCIAS - ESTADÍSTICA

DIRECTOR
RUBÉN DARÍO GUEVARA GONZALEZ, PH.D.
DOCTOR EN ESTADÍSTICA

LÍNEA DE INVESTIGACIÓN
CONTROL DE CALIDAD - DATOS FUNCIONALES



UNIVERSIDAD NACIONAL DE COLOMBIA
FACULTAD DE CIENCIAS
DEPARTAMENTO DE ESTADÍSTICA
BOGOTÁ, D.C.
2021

Título en español

Monitoreo de Perfiles No Lineales Multivariados usando un enfoque de Datos Funcionales

Title in English

Multivariate Nonlinear Profiles Monitoring Using a Functional Data Approach

Resumen: En este trabajo se presentan algunas propuestas para monitorear perfiles no lineales multivariados en fase II, usando métodos provenientes del análisis de datos funcionales. El desempeño de las cartas de control propuestas se evalúa usando simulaciones de Monte Carlo bajo diferentes escenarios. Para ilustrar el uso de la cartas propuestas se presentan ejemplos con datos reales.

Abstract: In this work, some proposals for the monitoring of multivariate non-linear profiles in phase II will be presented using statistical control charts, using an approach from the Functional Data Analysis. To evaluate the performance of the proposed charts, Monte Carlo simulations will be carried out under different scenarios. To illustrate the use of the proposed letters, examples with real data will be presented.

Palabras clave: Dato Funcional, Carta de Control, Perfiles no lineales, Profundidad multivariada funcional, Outlyingness funcional, Semidistancia de Mahalanobis Funcional.

Keywords: Functional data, Control chart, Non-linear profile, Functional multivariate Depth, Functional Outlyingness, Mahalanobis semidistance.

Nota de aceptación

Trabajo de tesis

Aprobado

“Mención Meritoria o Laureada”

Jurado

Jurado 1

Jurado

Jurado 2

Director

Rubén Darío Guevara

Bogotá, D.C., 2021

Dedicado a

Mi familia.

Agradecimientos

Un agradecimiento especial a mi director de tesis, el profesor Rubén Darío Guevara, por todas las enseñanzas y el constante apoyo en esta tarea. También a mi familia, quienes nunca me han dejado de apoyar y ser mi soporte. A la Universidad Nacional de Colombia, mi alma mater y el segundo hogar de la familia Espinosa Moreno. Finalmente a los estadísticos Tania López y Juan Camilo Bernal por sus consejos y apoyo, que resultaron muy significativos para el desarrollo de esta tesis.

Índice general

Índice general	I
Índice de tablas	III
Índice de figuras	V
Introducción	VII
1. Revisión de literatura	1
1.1. Datos funcionales univariados	1
1.1.1. Definición	1
1.1.2. Suavizamiento por mínimos cuadrados	3
1.1.3. Suavizamiento Spline	4
1.1.4. Análisis de componentes principales funcionales	6
1.1.5. Semidistancia funcional de Mahalanobis	9
1.1.6. Bootstrap suavizado	10
1.1.7. Carta de control basada en Sheu et al. [2013]	10
1.2. Datos funcionales multivariados	11
1.2.1. Definición	11
1.2.2. Suavizado	12
1.2.3. Profundidad para datos funcionales multivariados	13
1.2.3.1. Profundidad funcional multivariada de Claeskens	14
1.2.4. Medida outlyingness ajustada sobre datos funcionales multivariados .	15
1.2.4.1. Outlyingness univariada	15
1.2.4.2. Outlyingness ajustada	16
1.2.4.3. Outlyingness ajustada funcional	17
1.2.5. Componentes principales funcionales multivariadas	17

1.2.5.1. Componentes principales para datos funcionales multivariados MFPCA	17
1.2.5.2. Estimación de los componentes principales funcionales multivariados de X	19
1.2.6. Carta de control RSREWMA propuesta por Pan et al. [2019]	20
2. Propuestas metodológicas	23
2.1. Modelo	24
2.2. Enfoques propuestos	25
2.2.1. Carta basada en profundidad funcional multivariada (Carta MFHD)	25
2.2.2. Carta basada en la medida outlyingness ajustada funcional (Carta fAO)	27
2.2.3. Carta basada en la semidistancia de Mahalanobis funcional datos funcionales multivariados (Carta \mathbf{T}_{MF}^2)	29
3. Simulaciones	34
3.1. Evaluación del rendimiento de la carta de control	34
3.2. Comparación cartas univariadas vs multivariadas	35
3.3. Comparación cartas multivariadas	41
4. Aplicación a datos reales	51
Conclusiones	54
Trabajo futuro	55

Índice de tablas

3.1. Límites de control estimados para las cartas F , T_F^2 , T_{MF}^2 , fAO y MFHD usando bootstrap suavizado, con una tasa de falsas alarmas de 0.005 y su respectiva desviación estándar en paréntesis	37
3.2. Potencia promedio estimada de las cartas F , T_F^2 , T_{MF}^2 , fAO y MFHD con su respectiva desviación estándar en paréntesis, cuando el proceso es contaminado con el tipo de ruido R1	39
3.3. Potencia promedio estimada de las cartas F , T_F^2 , T_{MF}^2 , fAO y MFHD con su respectiva desviación estándar en paréntesis, cuando el proceso es contaminado con el tipo de ruido R2	39
3.4. Tiempos de ejecución de las estadísticas asociadas a las cartas F y T_F^2 para una curva estimada y T_{MF}^2 , fAO y MFHD para un vector de 3 curvas, donde cada una de ellas se encuentra observada en 100 puntos discretos del tiempo	41
3.5. Límites de control estimados para las cartas T_{MF}^2 , fAO y MFHD usando bootstrap suavizado con una tasa de falsas alarmas de 0.005 con su respectiva desviación estándar	43
3.6. Potencia de las cartas de control <i>SVM</i> (sin desviación estándar), T_{MF}^2 , fAO y MFHD, para los perfiles generados por las ecuaciones (3.7), con error que se distribuye normal $N_4(\mathbf{0}, \mathbf{\Sigma})$, $\mathbf{\Sigma}$ generada por la ecuación (3.8), con valores de correlación bajo ($\rho = 0.1$), medio ($\rho = 0.5$) y alto ($\rho = 0.9$), bajo el esquema de contaminación C1 con valores δ_{11} entre 0 y 1.5 y sus respectivas desviaciones estándar en paréntesis	44
3.7. Potencia de las cartas de control <i>SVM</i> (sin desviación estándar), T_{MF}^2 , fAO y MFHD, para los perfiles generados por las ecuaciones (3.7), con error que se distribuye normal $N_4(\mathbf{0}, \mathbf{\Sigma})$ y exponencial multivariada $MVE(\mathbf{\Sigma})$, $\mathbf{\Sigma}$ generada por la ecuación (3.8), bajo el esquema de contaminación C2 con valores δ_{12} entre 0 y 0.02 y sus respectivas desviaciones estándar en paréntesis	45
3.8. Potencia de las cartas de control <i>SVM</i> (sin desviación estándar), T_{MF}^2 , fAO y MFHD, para los perfiles generados por las ecuaciones (3.7), con error que se distribuye normal $N_4(\mathbf{0}, \mathbf{\Sigma})$, $\mathbf{\Sigma}$ generada por la ecuación (3.8) y correlación alta ($\rho = 0.9$) entre los procesos, bajo el esquema de contaminación C3 y sus respectivas desviaciones estándar en paréntesis	46

3.9. ARL promedio estimado de las cartas SVM , fAO , $MFHD$ y T_{MF}^2 con su respectiva desviación estándar en paréntesis, cuando el proceso es contaminado con el tipo de ruido F1	47
3.10. ARL promedio estimado de las cartas SVM , fAO , $MFHD$ y T_{MF}^2 con su respectiva desviación estándar en paréntesis, cuando el proceso es contaminado con el tipo de ruido F2	47
4.1. Límites de control para las cartas T_M^2F , fAO y $MFHD$, aplicadas al conjunto de datos de fluorescencia en el azúcar ajustados mediante suavizamiento spline	52

Índice de figuras

1.1.	Ajuste de un dato funcional por medio de bases de funciones	4
1.2.	Ejemplo de la recta que minimiza las distancias ortogonales de los puntos a ella	7
1.3.	Ejemplo de un conjunto de 6 datos funcionales multivariados, donde cada observación se compone de 3 funciones	12
2.1.	Observaciones de m perfiles no lineales multivariados	23
2.2.	Aproximación funcional de un conjunto de m perfiles no lineales multivariados	24
2.3.	Carta de control bajo el enfoque de la profundidad multivariada MFHD para monitorear perfiles no lineales multivariados	27
2.4.	Carta de control bajo el enfoque de outlyingness ajustada funcional para monitorear perfiles no lineales multivariados	28
2.5.	Carta de control bajo el enfoque de semidistancia T_{MF}^2 para monitorear perfiles no lineales multivariados	30
3.1.	Muestra de observaciones discretizadas bajo el esquema de la ecuacion (3.4) y en azul la función media μ_X	36
3.2.	Funciones estimadas para las observaciones y_i y sus dos primeras derivadas	36
3.3.	Muestra de 25 curvas estimadas para las observaciones y_i bajo control y el efecto de la contaminación de magnitud (a) y forma (b). Para las contaminaciones se presentan cambios pequeños, medianos y grandes, de acuerdo a los valores δ_1 y δ_2	38
3.4.	Corrida de una carta de control para las propuestas F , T_F^2 , T_{MF}^2 , fAO y MFHD en fase II para 25 observaciones nuevas, bajo la contaminación R2, con $\delta_2 = 0.2$	40
3.5.	Realización del perfil no lineal multivariado especificado por las ecuaciones (3.7). La fila superior corresponde a los dos primeros procesos del perfil, y la fila inferior corresponde a los dos últimos	48
3.6.	Suavizado del perfil no lineal multivariado especificado por las ecuaciones (3.7). La fila superior corresponde a los dos primeros procesos del perfil, y la fila inferior corresponde a los dos últimos	49

3.7. Contaminación para los dos primeros procesos del perfil no lineal multivariado, descritos en la ecuación 3.8, bajo los esquemas C1 con $\delta_1 = 0.6$ (a) y C2 con $\delta_2 = 0.01$ (b), respectivamente	50
3.8. Contaminación para los procesos los dos últimos procesos del perfil no lineal multivariado, descritos en la ecuación 3.8, bajo el esquema C3 con $\gamma_1 = 0.94$ (a) y $\gamma_3 = 1.06$ (b)	50
4.1. Muestra de 25 curvas estimadas para longitudes de onda 230 nm (a) y 290 nm (b) mediante suavizamiento spline, del conjunto de datos de fluorescencia en la producción de azúcar, en fase I	52
4.2. Cartas $T_M^2 F$, fAO y MFHD en fase II, para el conjunto de datos D_2	53

Introducción

En las últimas décadas el aumento en la recolección de datos ha sido de gran impacto en múltiples sectores tales como las industrias, los bancos, los gobiernos, etc. Este aumento en la cantidad de información ha permitido que algunas metodologías estadísticas, tales como el control de procesos, en el cual se verifica si un elemento posee las características de diseño con que fue planteado, vengán en constante desarrollo. Esto ha resultado fundamental tanto para la industria como para otras áreas de desarrollo como la medicina, la economía, la academia, entre otros, debido a la necesidad de adecuar y controlar los procesos para cumplir con estándares de calidad básicos y que permitan un desarrollo adecuado de las operaciones.

El control estadístico de procesos se compone de un conjunto de herramientas estadísticas que permiten realizar la evaluación de uno o múltiples procesos, tales como las cartas de control que serán descritas más adelante en detalle. El desarrollo de esta área de la estadística resulta relevante dado que, entre otras razones, permite monitorear procesos y evaluar estadísticamente si un producto está cumpliendo con los requerimientos esperados.

Los resultados de los procesos industriales, incluso cuando tienen comportamientos regulares, presentan variaciones con comportamientos aleatorios. Estas variaciones son causadas, en ocasiones, por causas comunes que no son controlables por los encargados de monitorear los procesos. Sin embargo, en otras ocasiones estas causas pueden ser totalmente controladas y así determinar que un proceso se encuentra bajo control [Qiu, 2013]. Cuando algunos elementos de un proceso no están bajo control, algunas características presentan alta variabilidad o variabilidad sistemática respecto a los requerimientos iniciales, y consecuentemente muchos productos no los cumplirán. Este tipo de variación se conoce con el nombre de causas asignables. Algunos ejemplos de estas son: los materiales, mal trabajo de los empleados, desajustes, etc. Desde la estadística se han propuesto las cartas de control como herramienta fundamental para evaluar estas variaciones en los procesos, las cuales se definen como un gráfico de las características de interés medidas versus el número de muestra o el tiempo en que cada observación fue medida. Típicamente, una carta de control está compuesta por una línea central (LC) que representa el promedio de los valores medidos y dos límites de control: límite de control superior (UCL) y límite de control inferior (LCL), que son construidos a partir de una serie de consideraciones estadísticas [Montgomery, 2007].

Se dirá que un proceso está bajo control si las únicas causas de variación son las causas comunes y la única forma de que estas cambien es cambiando completamente el proceso en sí mismo. De forma análoga diremos que un proceso está fuera de control si presenta causas asignables de variación [Qiu, 2013].

Generalmente, los procesos son monitoreados en dos fases: la primera, conocida como *fase I*, es la etapa en la cual se toma un conjunto de datos históricos y se analizan de forma retrospectiva, es decir, construyendo límites de control de prueba basados en la distribución de los datos y los cuales permitirán analizar datos que provienen del mismo proceso que se podrán monitorear de forma instantánea. En esta parte se entiende la variabilidad del proceso, se evalúa su estabilidad y se selecciona un modelo apropiado en control que proporcione un conjunto de estimadores de los parámetros que definen que el proceso se encuentre bajo control [Jones-Farmer et al., 2014].

La *fase II* empieza después de haber realizado un proceso de limpieza de datos que se encuentran bajo condiciones estables y representan un proceso bajo control. El objetivo es comparar nuevas muestras de una población, empleando las estadísticas obtenidas en la fase I con cada valor en la muestra de forma sucesiva y dibujándolas respecto a los límites de control establecidos anteriormente para determinar si se encuentran dentro de control o fuera de este [Montgomery, 2007].

Dependiendo de la dimensión de las características a observar las cartas de control pueden ser univariadas, en las cuales se monitorea solamente una característica, o multivariadas, donde se monitorean 2 o más características. Numerosos autores han descrito y explorado las cartas de control multivariadas dentro de las cuales se encuentran las cartas tipo Shewhart no paramétricas [Boone and Chakraborti, 2012], cartas basadas en modelos log-lineales [Qiu, 2013] y la carta T^2 de Hotelling [Montgomery, 2007], con la cual se pueden monitorear el vector de medias y/o la matriz de covarianza de un proceso. Otras cartas muy populares empleadas para monitorear procesos caracterizados por características de calidad multivariadas son la carta MCUSUM, entre las cuales se encuentran las versiones propuestas por Pignatiello and Runger [1990] y Crosier [1988], y la carta MEWMA, propuesta por Lowry et al. [1992] las cuales tienen supuestos de normalidad multivariada en los datos. Estas cartas son más sensibles para identificar cambios pequeños en el vector de medias respecto a cartas tales como la T^2 [Montgomery, 2007]. Dados los supuestos de normalidad multivariada sobre los conjuntos de datos para monitorear procesos de forma tradicional, también se han propuesto algunas cartas multivariadas no paramétricas basadas en rangos [Qiu, 2013].

En procesos más modernos, donde la cantidad de datos ha ido aumentando en la mayoría de las organizaciones, se han desarrollado modelos de control de procesos monitoreando perfiles, donde el objetivo es evaluar la calidad de un proceso mediante una relación entre una variable respuesta y un conjunto de variables explicativas, obteniendo así una relación funcional cuya estabilidad se evaluará en diferentes puntos del tiempo [Maleki et al., 2018]. Los perfiles pueden ser vistos como observaciones de procesos estocásticos cuya dimensión es infinita. Estos, a su vez, se pueden clasificar en dos tipos: lineales y no lineales. Un perfil es no lineal si por lo menos una derivada de la función media con respecto a los parámetros depende por lo menos de uno de los parámetros

[Schabenberger and Pierce, 2001].

Para monitorear los perfiles se han empleado algunas técnicas provenientes del análisis multivariado tales como Kang and Albin [2000], Noorossana et al. [2011] y Yeh et al. [2009], entre otros. Dentro de la estadística no paramétrica existen algunas propuestas tales como Williams et al. [2007], Zou et al. [2008], Shiau et al. [2009], Qiu et al. [2010], Zou et al. [2012], Chuang et al. [2013], Li et al. [2014]. Finalmente, bajo el análisis de datos funcionales algunos estudios tales como Sheu et al. [2013], Fassò et al. [2016], Paynabar et al. [2016] y Wang et al. [2018], entre los más relevantes.

El análisis de datos funcionales resulta muy importante al expandir los métodos de la estadística clásica, en particular, a objetos caracterizados por curvas con dominio en un espacio infinito-dimensional. Entre los autores más destacados en esta área encontramos a Ramsay and Silverman [2005], Ferraty and Vieu [2010], Horváth and Kokoszka [2012], Kokoszka and Reimherr [2017], entre otros.

La mayoría de los métodos estadísticos desarrollados para datos funcionales han sido para procesos univariados, es decir, conjuntos de datos donde se tienen N observaciones de un solo proceso estocástico en un intervalo compacto $[0, T]$. Sin embargo, también existen algunos métodos para el análisis de procesos funcionales multivariados, en los cuales se tienen N observaciones de p datos funcionales univariados, cada uno definido en un intervalo compacto, los cuales pueden ser iguales o diferentes. Estas observaciones pueden ser organizadas en una matriz de tamaño $N \times p$, en donde sus elementos son objetos funcionales (por ejemplo curvas). Uno de los métodos más conocidos es el análisis de componentes principales funcionales multivariadas (MFPCA) [Ramsay and Silverman, 2005], [Berrendero et al., 2011], [Jacques and Preda, 2014], y [Happ and Greven, 2018] en las cuales se busca reducir el número de dimensiones del proceso tal que la pérdida de información sea mínima y se pueda obtener un conjunto de valores que representen el proceso.

La mayoría de las propuestas realizadas para monitorear procesos caracterizados por perfiles basados en un enfoque de datos funcionales se enfocan en un solo proceso. Por ejemplo, Zhang et al. [2015] emplean una metodología de datos funcionales para ajustar perfiles multivariados en fase I, empleando componentes principales funcionales y con estos calculando una estadística T^2 para cada conjunto de datos.

En general, las cartas que han empleado datos funcionales multivariados los han enfocado en perfiles multi-canal, los cuales se definen como señales múltiples tomadas de diferentes fuentes [Grasso et al., 2014]. En algunos casos, el supuesto más fuerte es que los procesos son muy similares entre ellos, es decir, tienen correlaciones muy altas. Otra propuesta, realizada por Paynabar et al. [2016], monitorea perfiles no lineales multi-canal bajo el supuesto de que los perfiles tienen una estructura similar, es decir, las curvas de los perfiles exhiben patrones similares, y posteriormente se realiza reducción de dimensiones mediante componentes principales funcionales multivariadas, tomando las eigenfunciones y los eigenvalores asociados, para realizar una metodología de punto de cambio en fase I. En el trabajo realizado por Wang et al. [2018] se presenta una metodología para monitorear perfiles multi-canal empleando componentes principales

multivariadas funcionales de umbral, en las cuales primero se realiza una reducción del conjunto de datos mediante componentes principales funcionales multivariadas, obteniendo un conjunto de características tanto para las curvas dentro de control como para las curvas fuera de control, teniendo en cuenta que en otras propuestas similares solamente se trabajan componentes principales para las curvas bajo control conllevando a retenciones de pocas componentes principales que garanticen alta información solamente bajo control; luego se emplea una metodología de punto de cambio de forma similar a Paynabar et al. [2016]. También se tiene la carta presentada por Pan et al. [2019], en la cual se ajustan modelos de regresión no paramétrica en fase II y la carta presentada por Zhang et al. [2018], en la cual se ajustan perfiles multivariados mediante datos funcionales multivariados débilmente correlacionados a partir del desarrollo de las componentes principales funcionales *Sparse multichannel*, en las que se combinan los conceptos de componentes principales funcionales y de regresión lineal multivariada con penalización LASSO en los parámetros. Recientemente, la metodología propuesta por Wang and Tsung [2020] desarrolla una definición nueva de componentes principales funcionales multivariadas *escasas jerárquicas*, las cuales realizan un modelamiento conjunto de los perfiles por etapas para identificar qué variables son las más informativas en cada eigenvector. Esto con el objetivo de interpretar mejor la reducción de dimensiones a partir de reparametrizaciones y reformulaciones de las componentes tradicionales y optimizando cuando se tienen datos en altas dimensiones. Atashgar and Zargarabadi [2017] proponen una metodología aplicada al control y monitoreo de variables asociadas a la manufactura de plástico empleando perfiles y el nivel de contribución de cada variable empleando la estadística λ de Wilks en fase I y Jahani et al. [2018] proponen una metodología de monitoreo de perfiles multivariados empleando procesos Gaussianos multivariados, los cuales son ajustados para modelar un conjunto de datos históricos bajo control considerando tanto la correlación entre procesos, como la correlación dentro de cada uno.

De forma complementaria a estos trabajos, en esta tesis se realiza un aporte adicional a esta área de conocimiento empleando técnicas de datos funcionales multivariados, independientemente si son perfiles multichannel o no, tales como profundidades, outlyingness y semidistancia de Mahalanobis funcional aplicando componentes principales funcionales desarrolladas por Happ and Greven [2018]. En las metodologías presentadas, las observaciones discretizadas son suavizadas mediante suavizamiento spline. En los resultados reportados por Ramsay and Silverman [2005], se denota que ajustar expansiones de base por mínimos cuadrados ordinarios implica control discontinuo sobre el grado de suavizado, el cual puede ser mejorado por el suavizamiento spline, en el cual se realiza una penalización de acuerdo al nivel de rugosidad de la curva, definida más adelante. También se emplea métodos de remuestreo (bootstrap) para datos funcionales, con el objetivo de calcular los límites de control sin supuestos distribucionales.

El trabajo se estructura de la siguiente manera: en el primer capítulo se realiza una revisión de la literatura referente a cartas de control y datos funcionales multivariados, con el objetivo de examinar el estado del arte y evaluar los avances en el tema. Luego, en el segundo capítulo, se describe la metodología propuesta para el desarrollo de las cartas de control para perfiles multivariados en fase II, empleando técnicas del análisis de datos funcionales. En el capítulo 3 se reportan los resultados de las simulaciones realizadas con base en el modelo propuesto en el capítulo 2. En el capítulo 4 se realiza una aplicación de los modelos propuestos sobre un conjunto de datos reales. Para finalizar se presenta

una serie de conclusiones con base en los resultados obtenidos de las simulaciones, trabajos futuros a desarrollar y se presenta la bibliografía que permitió la realización del mismo.

Revisión de literatura

En la primera sección de esta revisión se presentarán los conceptos claves respecto al suavizamiento de datos funcionales multivariados y, de forma más específica, los trabajos realizados referentes a profundidades Cuevas et al. [2007], [Claeskens et al., 2014] y reducción de dimensiones mediante componentes principales funcionales multivariados [Jacques and Preda, 2014], [Górecki et al., 2016]. También se presentarán algunas de las cartas de control que se han propuesto recientemente para monitoreo de perfiles no lineales univariados [Sheu et al., 2013] y multivariados [Pan et al., 2019].

1.1. Datos funcionales univariados

1.1.1. Definición

En diferentes campos de las ciencias, los datos observados se pueden asociar a observaciones de una función definida sobre un intervalo en algún conjunto tal como los números reales, entre otros. El análisis de datos funcionales busca el modelamiento y correcto análisis de estos conjuntos de datos de acuerdo a la naturaleza del problema de estudio. En la práctica los valores de las funciones son observados en un número finito de puntos discretos donde, en múltiples ocasiones, se encuentran menos individuos que puntos sobre los cuales se evalúan, razón por la cual los métodos clásicos del análisis multivariado no se podrían aplicar [Galeano et al., 2015]. Para el análisis de este tipo de datos se ha desarrollado una serie de metodologías y técnicas que se pueden profundizar en Ramsay and Silverman [2005], Horváth and Kokoszka [2012] y Kokoszka and Reimherr [2017], entre otros.

A continuación se presenta un resumen de algunos de los métodos que se han desarrollado y que se emplearon para el desarrollo de la tesis.

El espacio $L_2 = L_2([a, b])$ es el conjunto de funciones reales x medibles, definidas sobre el intervalo $[a, b]$ con $a < b$, $a, b \in \mathbb{R}$, que satisfacen la condición $\int_a^b x^2(t)dt < \infty$. El espacio L_2 es un espacio separable de Hilbert [Horváth and Kokoszka, 2012] con el

siguiente producto punto:

$$\langle x, y \rangle = \int_a^b x(t)y(t)dt. \quad (1.1)$$

Sea X una curva aleatoria definida en un intervalo compacto $I = [a, b]$. Se dice que X es integrable si $E[\|X\|] = E[\int_I X^2(t)dt]^{1/2} < \infty$. Si X es integrable, existe una única función $\mu_X \in L_2$ tal que $E\langle y, X \rangle = \langle y, \mu_X \rangle$ para cualquier $y \in L_2$. De aquí sigue que $\mu_X(t) = E[X(t)]$, $t \in I$, es decir, la función media del proceso X .

De forma similar, se dice que X es cuadrado integrable si:

$$E[\|X\|^2] = E \int_I X^2(t)dt < \infty. \quad (1.2)$$

El espacio L_2 se define como el conjunto de todas las funciones cuadrado integrables.

Si X es una función aleatoria cuadrado integrable y $E[X] = 0$ entonces el operador de covarianza se define así [Horváth and Kokoszka, 2012]:

$$C(y) = E[\langle X, y \rangle X], \quad y \in L_2. \quad (1.3)$$

Cuando el proceso X no se encuentre centrado, se centra restándole su valor medio μ_X .

Es fácil observar que, para cualquier par de puntos $t, s \in I$ se cumple

$$C(y)(t) = \int_I c(t, s)y(s)ds, \quad \text{donde } c(t, s) = E[X(t)X(s)]. \quad (1.4)$$

Este operador C es simétrico, es decir, $c(t, s) = c(s, t)$ y

$$E \left[(X(t)y(t)dt)^2 \right] \geq 0, \quad (1.5)$$

es decir, es un operador definido positivo. Así, C tiene un conjunto $\lambda_1, \lambda_2, \dots$ de eigenvalores no negativos (ver Horváth and Kokoszka [2012]) que satisfacen:

$$\sum_{j=1}^{\infty} \lambda_j < \infty, \quad (1.6)$$

y un conjunto de funciones propias ortogonales ψ_1, ψ_2, \dots que se pueden normalizar y forman una base en L_2 .

Bajo estas condiciones, cada curva $X \in L_2$ se puede descomponer de la siguiente manera:

$$X = \mu_X + \sum_{k=1}^{\infty} \theta_k \psi_k, \quad (1.7)$$

la cual es conocida como la descomposición de *Karhunen-Loève*, donde $\theta_k = \langle X - \mu_X, \psi_k \rangle$ son los scores de las observaciones funcionales, es decir, las proyecciones de las observaciones funcionales sobre las funciones propias. Cuando se selecciona un número K de componentes a retener, la descomposición de la curva queda aproximada de forma finita

de la siguiente manera:

$$X \approx \mu_X + \sum_{k=1}^K \theta_k \psi_k. \quad (1.8)$$

1.1.2. Suavizamiento por mínimos cuadrados

Sean y_i , $i = 1, \dots, n$ un conjunto de observaciones discretizadas de un dato funcional X en un intervalo compacto real $I = [a, b]$ con $a < b$, $a, b \in \mathbb{R}$. El objetivo es ajustar una curva a las observaciones discretizadas, que siga el modelo $y_j = x(t_j) + \epsilon_j$ donde $t_j \in I$ y ϵ_j es un error aleatorio de media 0.

Se empleará una expansión en bases de funciones $\phi_1, \phi_2, \dots, \phi_K$, para x [Ramsay and Silverman, 2005].

$$x(t) = \sum_{k=1}^K c_k \phi_k(t) = \mathbf{c}^T \boldsymbol{\phi}. \quad (1.9)$$

El vector \mathbf{c} de longitud K contiene los coeficientes c_k . Se define $\boldsymbol{\Phi}$ la matriz de tamaño $n \times K$ que contiene los valores $\phi_k(t_j)$:

$$\boldsymbol{\Phi} = \begin{bmatrix} \phi_1(t_1) & \phi_2(t_1) & \dots & \phi_K(t_1) \\ \phi_1(t_2) & \phi_2(t_2) & \dots & \phi_K(t_2) \\ \vdots & \vdots & \ddots & \vdots \\ \phi_1(t_n) & \phi_2(t_n) & \dots & \phi_K(t_n) \end{bmatrix}. \quad (1.10)$$

Empleando el método de los mínimos cuadrados ordinarios se define la siguiente función a minimizar:

$$\text{SMSSE}(\mathbf{y}|\mathbf{c}) = \sum_{j=1}^n \left[y_j - \sum_{k=1}^K c_k \phi_k(t_j) \right]^2, \quad (1.11)$$

donde $\mathbf{y} = (y_1, y_2, \dots, y_n)^T$. La expresión en la ecuación 1.11 queda en la siguiente forma matricial:

$$\text{SMSSE}(\mathbf{y}|\mathbf{c}) = (\mathbf{y} - \boldsymbol{\Phi}\mathbf{c})^T (\mathbf{y} - \boldsymbol{\Phi}\mathbf{c}). \quad (1.12)$$

Luego, para minimizar esta suma se deriva con respecto a \mathbf{c} , obteniendo

$$2\boldsymbol{\Phi}\boldsymbol{\Phi}^T \mathbf{c} - 2\boldsymbol{\Phi}^T \mathbf{y} = 0. \quad (1.13)$$

Resolviendo para \mathbf{c} se obtiene la estimación $\hat{\mathbf{c}}$ como sigue

$$\hat{\mathbf{c}} = (\boldsymbol{\Phi}^T \boldsymbol{\Phi})^{-1} \boldsymbol{\Phi}^T \mathbf{y}. \quad (1.14)$$

Luego el vector de los valores ajustados será

$$\hat{\mathbf{y}} = \boldsymbol{\Phi}\hat{\mathbf{c}} = \boldsymbol{\Phi}(\boldsymbol{\Phi}^T \boldsymbol{\Phi})^{-1} \boldsymbol{\Phi}^T \mathbf{y}. \quad (1.15)$$

En Ramsay and Silverman [2005] se recomienda que se emplee el método de los mínimos cuadrados cuando los errores ϵ_j son independientes idénticamente distribuidos con media 0 y varianza $\sigma^2 \in \mathbb{R}^+$.

Cuando se emplea el método de los mínimos cuadrados ponderados, entonces se define la siguiente función a minimizar:

$$\text{SMSSE}_m(\mathbf{y}|\mathbf{c}) = (\mathbf{y} - \Phi\mathbf{c})^T \mathbf{W}(\mathbf{y} - \Phi\mathbf{c}), \quad (1.16)$$

obteniendo como solución:

$$\hat{\mathbf{c}} = (\Phi^T \mathbf{W} \Phi)^{-1} \Phi^T \mathbf{W}^T \mathbf{y}, \quad (1.17)$$

donde Φ es una matriz de tamaño $n \times K$ que contiene los valores de las K bases de funciones en los n puntos muestrales, \mathbf{W} es una matriz de pesos para evaluar posibles estructuras de covarianza entre los residuales, y \mathbf{y} es el vector de datos discretos a ser suavizados [Ramsay and Silverman, 2005]. Luego, el vector de valores ajustados es

$$\hat{\mathbf{y}} = \Phi(\Phi^T \mathbf{W} \Phi)^{-1} \Phi^T \mathbf{W} \mathbf{y} = S_\phi \mathbf{y}, \quad (1.18)$$

donde S_ϕ es el operador de proyección [Ramsay and Silverman, 2005]:

$$S_\phi = \Phi(\Phi^T \mathbf{W} \Phi)^{-1} \Phi^T \mathbf{W}, \quad (1.19)$$

correspondiente al sistema de bases ϕ . En la Figura 1.1 se puede observar el dato funcional

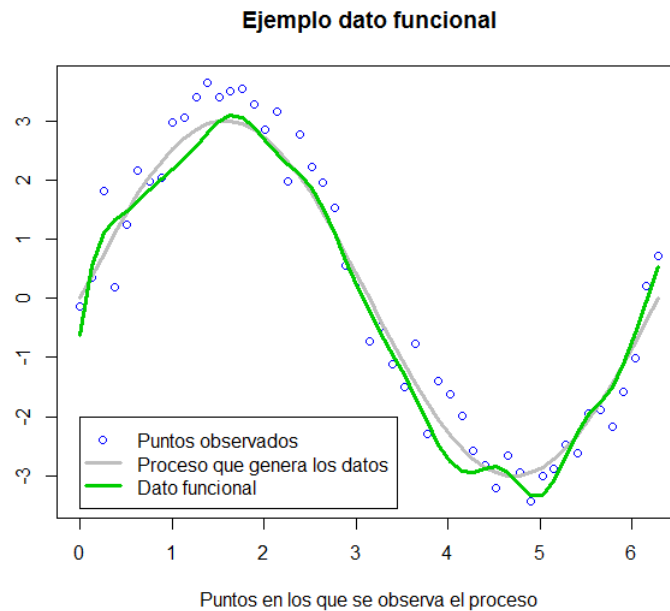


FIGURA 1.1. Ajuste de un dato funcional por medio de bases de funciones

ajustado a partir de un conjunto de observaciones discretizadas.

1.1.3. Suavizamiento Spline

En esta subsección se describirá el suavizamiento spline, el cual emplea la *regularización* de las curvas con el objetivo de penalizar el suavizamiento en la medida que las curvas sean muy ruidosas [Ramsay and Silverman, 2005].

Para cuantificar el nivel de rugosidad de una curva $x(t)$ para $t \in I = [a, b]$ con $a < b$, $a, b \in \mathbb{R}$ se emplea el cuadrado de la segunda derivada $[D^2x(t)]^2$, el cual se conoce como la curvatura de x . Se denominará

$$\text{PEN}_2(x) = \int [D^2x(s)]^2 ds, \quad (1.20)$$

a la función que mide la penalidad por rugosidad de una curva con base en la segunda derivada.

En general, se generaliza el concepto de penalidad por rugosidad de la siguiente manera [Ramsay and Silverman, 2005]:

$$\text{PEN}_m(x) = \int [D^m x(s)]^2 ds, \quad (1.21)$$

donde $D^m x$ es la derivada de orden m de la función x .

Sea $x(\mathbf{t})$ el vector que resulta de evaluar la función x en el vector de argumentos \mathbf{t} . Se define la suma de residuales penalizada, de forma similar a la sección anterior, como

$$\text{PENSSE}_\lambda(x|\mathbf{y}) = [\mathbf{y} - x(\mathbf{t})]^T \mathbf{W} [\mathbf{y} - x(\mathbf{t})] + \lambda \times \text{PEN}_2(x). \quad (1.22)$$

Luego la estimación de la función será obtenida encontrando x que minimice $\text{PENSSE}_\lambda(x)$ sobre el espacio de funciones x para el cual $\text{PEN}_2(x)$ esté definido [Ramsay and Silverman, 2005].

El parámetro λ es un parámetro de suavizado, específicamente, es la tasa que mide el tipo de cambio entre el ajuste a los datos, medido por la suma residual de cuadrados en el primer término, y la variabilidad de la función x , cuantificada por $\text{PEN}_2(x)$ en el segundo término [Ramsay and Silverman, 2005]. A medida que λ sea más grande, las funciones que no son lineales deben incluir un valor de penalización más alto a través del término $\text{PEN}_2(x)$. Por otro lado, mientras λ tienda a 0, la tendencia de la curva será más variable ya que la penalización será pequeña, llevando al método a realizar una interpolación más que un suavizado de los datos. Dentro de los métodos más empleados para calcular el valor óptimo de λ se encuentran la validación cruzada y la validación cruzada generalizada [Ramsay and Silverman, 2005].

La computación de los splines se realiza de la siguiente manera:

- Recordemos que la curva se reescribe en términos de las bases de funciones

$$x(t) = \sum_{k=1}^K c_k \phi_k(t) = \mathbf{c}^T \boldsymbol{\phi}(t) = \boldsymbol{\phi}(t)^T \mathbf{c}, \quad (1.23)$$

donde \mathbf{c} es el vector de coeficientes y $\boldsymbol{\phi}$ es el vector de bases de funciones. El factor de penalidad $\text{PEN}_m(x)$ se puede reexpresar de forma matricial de la siguiente manera

[Ramsay and Silverman, 2005]:

$$\begin{aligned}
\text{PEN}_m(x) &= \int [D^m x(s)]^2 ds \\
&= \int [D^m \mathbf{c}^T \boldsymbol{\phi}(s)]^2 ds \\
&= \int \mathbf{c}^T D^m \boldsymbol{\phi}(s) D^m \boldsymbol{\phi}^T(s) \mathbf{c} ds \\
&= \mathbf{c}^T \left[\int D^m \boldsymbol{\phi}(s) D^m \boldsymbol{\phi}^T(s) ds \right] \mathbf{c} \\
&= \mathbf{c}^T \mathbf{R} \mathbf{c},
\end{aligned} \tag{1.24}$$

donde

$$\mathbf{R} = \int D^m \boldsymbol{\phi}(s) D^m \boldsymbol{\phi}^T(s) ds. \tag{1.25}$$

Adicionando el término de error y el de penalización, multiplicado por λ , se tiene

$$\text{PENSSE}_m(\mathbf{y}|\mathbf{c}) = (\mathbf{y} - \boldsymbol{\Phi} \mathbf{c})^T \mathbf{W} (\mathbf{y} - \boldsymbol{\Phi} \mathbf{c}) + \lambda \mathbf{c}^T \mathbf{R} \mathbf{c}, \tag{1.26}$$

obteniendo las siguientes expresiones tanto para el vector de parámetros $\hat{\mathbf{c}}$, como para los datos ajustados $\hat{\mathbf{y}}$, respectivamente:

$$\hat{\mathbf{c}} = (\boldsymbol{\Phi}^T \mathbf{W} \boldsymbol{\Phi} + \lambda \mathbf{R})^{-1} \boldsymbol{\Phi}^T \mathbf{W} \mathbf{y} \tag{1.27}$$

$$\hat{\mathbf{y}} = \boldsymbol{\Phi} (\boldsymbol{\Phi}^T \mathbf{W} \boldsymbol{\Phi} + \lambda \mathbf{R})^{-1} \boldsymbol{\Phi}^T \mathbf{W} \mathbf{y} = \mathbf{S}_{\phi, \lambda} \mathbf{y}. \tag{1.28}$$

1.1.4. Análisis de componentes principales funcionales

El análisis de componentes principales es fundamental en el análisis multivariado de datos, dado que facilita la visualización y comprensión de la estructura de covarianza de un conjunto de datos mediante la reducción de las dimensiones de conjuntos de datos. El objetivo es que, dadas n observaciones de p variables, se analiza si es posible representar adecuadamente esta información con un número menor de variables construidas como combinaciones lineales de las originales [Peña, 2002]. En este tipo de análisis se puede evidenciar gráficamente la relación existente entre observaciones (individuos) y características (variables), entre otros. En datos funcionales resulta muy importante este análisis, dado que se pasa de trabajar en un espacio infinito dimensional de las curvas a trabajar en un espacio finito, en el cual se retiene un porcentaje alto de la variabilidad que garantiza una representación adecuada de los datos [Peña, 2002].

En componentes principales sobre datos multivariados el problema que se desea resolver es cómo encontrar un espacio de dimensión más reducida que represente adecuadamente los datos. Para un conjunto de n observaciones de p variables en una matriz $\mathbf{X}_{n \times p}$, se desea encontrar un subespacio de dimensión menor que p tal que al proyectar sobre él los puntos conserven su estructura con la menor distorsión posible. Si consideramos un punto \mathbf{x}_i y una dirección $\mathbf{a}_1 = (a_{11}, \dots, a_{1p})^T$, definida por un vector \mathbf{a}_1 normalizado, la proyección

del punto \mathbf{x}_i sobre esta dirección es el escalar:

$$z_i = a_{11}x_{i1} + \dots + a_{1p}x_{ip} = \mathbf{a}_1 \mathbf{x}_i \quad (1.29)$$

y el vector que representa esta proyección será $z_i \mathbf{a}_1$. Llamando r_i a la distancia entre el punto \mathbf{x}_i , y su proyección sobre la dirección \mathbf{a}_1 , este criterio implica:

$$\text{minimizar } \sum_{i=1}^n r_i^2 = \sum_{i=1}^n |\mathbf{x}_i - z_i \mathbf{a}_1|^2, \quad (1.30)$$

donde $|u|$ es la norma euclídea o módulo del vector u [Peña, 2002].

El caso para dos variables x_1, x_2 se visualiza en la gráfica 1.2, en la cual se evidencia una componente principal, calculada como la recta que minimiza las distancias ortogonales de un conjunto de puntos a ella (tomada de Peña [2002]). De este proceso de optimización

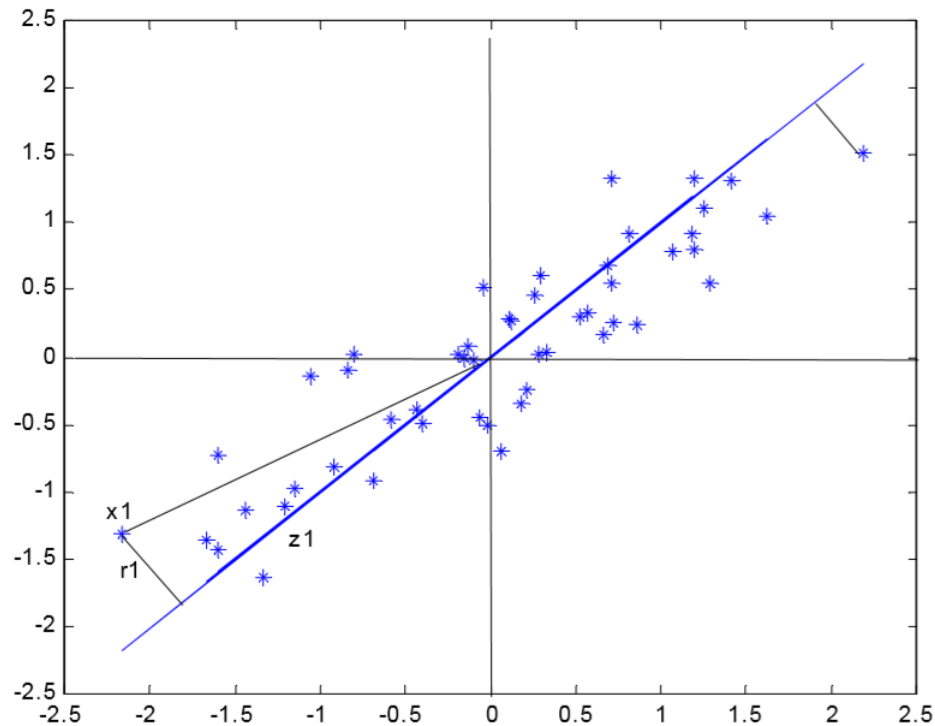


FIGURA 1.2. Ejemplo de la recta que minimiza las distancias ortogonales de los puntos a ella

se obtiene que los valores propios de la matriz $\mathbf{S} = 1/n(\mathbf{X}^T \mathbf{X})$ (matriz de varianzas y covarianzas de las observaciones), y sus vectores propios asociados, son los que permiten minimizar estas distancias y reducir las dimensiones.

Esta misma idea es llevada de datos multivariados a datos funcionales de forma análoga. Definamos X como una función definida sobre un intervalo compacto $I = [a, b]$ con $a < b$, $a, b \in \mathbb{R}$ con función media μ_X y operador de covarianza Γ_X . Si X es una función cuadrado integrable, entonces este operador es compacto (Mas (2007) mencionado en

Galeano et al. [2015]). Existe un conjunto de eigenvalores y funciones propias tal que

$$\Gamma_X(\psi_k) = \lambda_k \psi_k, \quad k = 1, 2, \dots \quad (1.31)$$

El objetivo es encontrar una función de pesos ϕ_1 tal que se maximice la varianza de la proyección

$$\theta_1 = \langle \phi_1, X - \mu_X \rangle, \quad (1.32)$$

sujeto a la restricción

$$\|\phi_1\|^2 = \int_I \phi_1(t)^2 dt = 1. \quad (1.33)$$

La solución a este problema de optimización es $\phi_1 = \Psi_1$, es decir, que la función de pesos que maximiza la varianza de la proyección sobre la curva centrada es la función propia asociada al mayor eigenvalor λ_1 . Una vez se reemplaza ϕ_1 por ψ_1 en la ecuación (1.32) se obtienen los scores de la función en la siguiente ecuación.

$$\theta_k = \langle X - \mu_X, \psi_k \rangle, \quad k = 1, 2, \dots, \quad (1.34)$$

donde se obtiene que la varianza de θ_1 es igual a λ_1 [Galeano et al., 2015].

En la práctica, donde se tiene una muestra aleatoria de n curvas x_i , $i = 1, 2, \dots, n$, se estiman las componentes principales empezando con una estimación de la función media del proceso en la siguiente ecuación.

$$\hat{\mu}_X = \frac{1}{n} \sum_{i=1}^n x_i, \quad (1.35)$$

y con una estimación del operador de covarianza en la siguiente ecuación [Galeano et al., 2015].

$$\hat{\Gamma}_X(\eta) = \frac{1}{n-1} \sum_{i=1}^n \langle x_i - \hat{\mu}_X, \eta \rangle (x_i - \hat{\mu}_X), \quad (1.36)$$

para cualquier función $\eta \in L_2(I)$.

Las funciones propias y los eigenvalores del operador Γ_X se estiman con base en $\hat{\Gamma}_X$ obteniendo el conjunto $\hat{\psi}_1, \hat{\psi}_2, \dots$ de funciones propias estimadas y $\hat{\lambda}_1, \hat{\lambda}_2, \dots$ los eigenvalores estimados.

Finalmente los scores estimados son

$$\hat{\theta}_{ik} = \langle x_i - \hat{\mu}_X, \hat{\psi}_k \rangle. \quad (1.37)$$

Este algoritmo se repite de acuerdo al número de componentes K que el investigador haya definido a priori de acuerdo a algún método de selección. Dentro de los métodos para seleccionar el número de componentes principales a retener se encuentra el porcentaje de varianza retenido basado en el *scree plot*, el cual es un gráfico de barras donde se grafican los eigenvalores estimados, de mayor a menor, evaluando el porcentaje de varianza que aporta cada uno. También se encuentra el criterio del porcentaje de varianza acumulada con cada componente principal funcional [Ramsay and Silverman, 2005]. En los estudios

realizados por Happ and Greven [2018], se enfatiza en la importancia de seleccionar de forma correcta el número de componentes principales a retener. Para determinar este valor no existe un criterio explícito, sin embargo, los métodos hasta ahora expuestos se basan en el porcentaje de varianza que se va acumulando en la medida que vaya aumentando el valor. Estos criterios pueden variar dependiendo el área donde se aplique y la naturaleza del estudio.

1.1.5. Semidistancia funcional de Mahalanobis

En esta subsección se realiza una descripción acerca del uso de la semidistancia de Mahalanobis propuesta por Galeano et al. [2015], con el objetivo de monitorear perfiles no lineales univariados y más adelante perfiles no lineales multivariados.

Sea X un dato funcional univariado definido en un intervalo compacto $I = [a, b]$ con $a < b$, $a, b \in \mathbb{R}$. Existe un conjunto de eigenvalores $\lambda_1, \lambda_2, \dots$ y un conjunto de funciones propias ψ_1, ψ_2, \dots tal que se aplica la descomposición de Karhunen-Loève a cada curva:

$$X = \mu_X + \sum_{k=1}^{\infty} \theta_k \psi_k, \quad (1.38)$$

donde μ_X es la media del proceso y $\theta_k = \langle X - \mu_X, \psi_k \rangle$ son los scores de la curva X , es decir, las proyecciones de esta curva sobre las funciones propias.

Luego se define la semidistancia funcional de Mahalanobis, entre una curva X y la media del proceso μ_X , como en la ecuación 1.39.

$$d_{FM}(X, \mu_X) = \left(\sum_{k=1}^K \omega_k^2 \right)^{1/2}, \quad \omega_k = \theta_k / \lambda_k^{1/2} \text{ (para } k = 1, \dots, K), \quad (1.39)$$

donde θ_k son los scores, proyecciones de las curvas sobre las funciones propias del operador de covarianza y λ_k son los eigenvalores del operador de covarianza y también representan la varianza de los scores. Estos se obtienen al realizar un ACP funcional donde el número de componentes retenidas es K , el cual, para esta Tesis, se selecciona de acuerdo a criterios de diferencia en la proporción de varianza acumulada, con los cuales se puede garantizar que se tiene poca pérdida de información en la reducción de dimensión.

Esta es una semidistancia funcional, dado que si la distancia entre dos curvas es cero, esto no implica que estas sean iguales. Esto no afecta la comparación entre las curvas y la función media, pues si dos curvas tienen la misma distancia, pero no son exactamente la misma, serán similares (respecto al proceso estocástico que las genera, bajo la descomposición de Karhunen-Loève) y ambas serán asignadas igualmente, bien sea bajo control o fuera de control.

Para esta Tesis se empleará la estadística T^2 funcional para datos univariados, para una curva X respecto a su vector de medias μ_X , como sigue:

$$T_F^2(X, \mu_X) = d_{FM}^2(X, \mu_X) = \sum_{k=1}^K \omega_k^2, \quad (1.40)$$

donde $\omega_k^2 = \theta_k^2/\lambda_k$ para $k = 1, \dots, K$ y K es el número de componentes principales a seleccionar.

1.1.6. Bootstrap suavizado

El bootstrap suavizado es una técnica empleada para obtener la distribución de un estadístico calculado sobre un conjunto de datos funcionales. Esta técnica resulta muy útil cuando no se tienen supuestos sobre la distribución del conjunto de curvas o del proceso del cual provienen las observaciones y también es ideal para evitar la aparición de observaciones repetidas.

De acuerdo a Cuevas et al. [2006] se define el bootstrap suavizado como el siguiente conjunto de pasos:

1. Sean X_1, X_2, \dots, X_n un conjunto de n curvas observadas y $T = T(X_1, \dots, X_n)$ la estadística de interés que se calculará sobre la muestra.

2. Se calcula la muestra X_1^0, \dots, X_n^0 a partir de un remuestreo usando el siguiente procedimiento: se define $X_i^0 = X_i^* + Z$, donde X_i^* es una muestra con reemplazamiento sobre el conjunto $\{X_1, X_2, \dots, X_n\}$, Z es un vector aleatorio que se distribuye normal con vector de medias $\mathbf{0}$ y matriz de varianzas y covarianzas $\gamma\Sigma_x$, donde Σ_x es la matriz de covarianza de $X_1^*, X_2^*, \dots, X_n^*$ y γ es el parámetro de suavizado, el cual Cuevas et al. [2006] recomiendan sea $\gamma = 0.05$.

3. Sea $T^b = T(X_1^b, \dots, X_n^b)$ el estimador usando la b -ésima muestra bootstrap.

4. Una vez calculada la estadística de interés T^b se repite el proceso B veces tomando finalmente el promedio como el estimador del parámetro de interés $\bar{T} = \sum_{b=1}^B \frac{T^b}{B}$. Para esta Tesis se emplean valores $B = 500$ replicaciones del método.

1.1.7. Carta de control basada en Sheu et al. [2013]

Esta carta se emplea para monitorear perfiles ajustados mediante datos funcionales univariados en fase II. Sea x_1, x_2, \dots, x_n funciones aleatorias independientes, tomadas de un conjunto histórico de datos con la misma distribución. El objetivo de Sheu et al. [2013] es construir una carta de control para estos datos. Si se observan puntos discretos de x_1, x_2, \dots, x_n en el tiempo, se emplean técnicas de análisis de datos funcionales para reconstruir la forma suavizada de estos. Si los datos funcionales observados son y_1, y_2, \dots, y_n dados por

$$y_i(t) = x_i(t) + \epsilon_i(t), \quad i = 1, 2, \dots, n, \quad t \in [a, b], \quad (1.41)$$

donde ϵ_i representa el error del equipo de medida y es independiente a x_i . Generalmente se asume que $\text{Var}(\epsilon_i(t)) = \sigma^2$ para cada t , con σ^2 desconocido. Por lo tanto, el primer paso para construir la carta de control basada en datos funcionales es estimar x_i individualmente, empleando suavizamiento spline. Posteriormente se emplea la estimación de x_i para construir la carta [Sheu et al., 2013].

Suponga que x es la curva observada, μ es la curva media y σ es la función desviación estándar. Sin pérdida de generalidad, primero se considera la situación en la cual el aumento es fijo. Se emplea la estadística

$$\frac{x - \mu}{\sigma}, \quad (1.42)$$

para evaluar si la curva x está bajo control o fuera de control punto a punto. Para monitorear los procesos, se emplea la estadística F, definida a continuación [Sheu et al., 2013]:

$$F = \int_{t_{min}}^{t_{max}} \left(\frac{x(t) - \mu(t)}{\sigma(t)} \right)^2 dt. \quad (1.43)$$

El objetivo del monitoreo con esta carta es encontrar en una primera fase los valores bajo control de μ y σ , los cuales se comparan con las nuevas curvas que van llegando en fase II. Para el desarrollo de esta tesis, esta carta será denominada la carta F.

1.2. Datos funcionales multivariados

1.2.1. Definición

De la misma forma en que fueron definidos los datos funcionales univariados, como datos que se pueden asociar a observaciones de una función definida sobre un intervalo en algún conjunto, los datos funcionales multivariados buscan el modelamiento y correcto análisis de estos conjuntos de datos, de acuerdo a la naturaleza del problema de estudio donde existe más de una característica medida sobre la misma observación, es decir, en este caso una observación funcional será un vector finito dimensional cuyos elementos son funciones, que podrán ser vistas como trayectorias de un proceso estocástico definido sobre un espacio de funciones infinito dimensionales [Berrendero et al., 2011]. La mayoría de los métodos que se han desarrollado para datos funcionales univariados han sido extendidos a datos funcionales multivariados, permitiendo así la inclusión de mayor información en los análisis en múltiples áreas del conocimiento. A continuación se presentará un resumen de algunos de los métodos que se han desarrollado y que se emplearon para el desarrollo de la tesis.

En la Figura 1.3, tomada de Berrendero et al. [2011], se puede observar visualmente un dato funcional multivariado. Sea $\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_n$ una muestra independiente idénticamente distribuida de un dato funcional multivariado \mathbf{X} . La observación de $\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_n$ provee un conjunto de datos funcional multivariado. Se considerarán los siguientes supuestos:

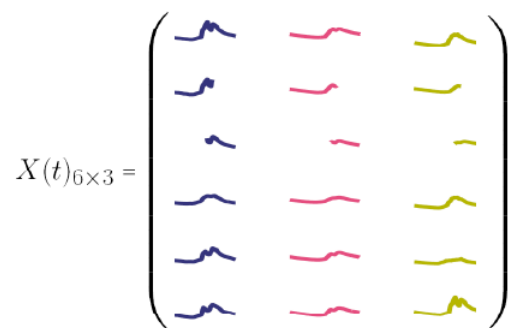


FIGURA 1.3. Ejemplo de un conjunto de 6 datos funcionales multivariados, donde cada observación se compone de 3 funciones

1. $\mathbf{X} = (X^1, \dots, X^p)^T$ es un proceso estocástico continuo en $L_2(I)$ [Jacques and Preda, 2014], es decir:

$$\forall t \in [a, b], \lim_{h \rightarrow 0} E [\|\mathbf{X}(t+h) - \mathbf{X}(t)\|^2] = \lim_{h \rightarrow 0} \int_a^b \sum_{l=1}^p E [(X^l(t+h) - X^l(t))^2] dt = 0. \quad (1.44)$$

Dado que \mathbf{X} es continuo en $L_2(I)$, cada una de sus funciones X^l , $l = 1, \dots, p$ también lo es. Denotemos $\mu_l = \{\mu_l = E[X_l]\}_{t \in [a, b]}$ a la función de media del proceso X^l para $1 \leq l \leq p$ y

$$\boldsymbol{\mu} = (\mu_1, \mu_2, \dots, \mu_p)^T = E[\mathbf{X}], \quad (1.45)$$

la función de media de \mathbf{X} .

El operador de covarianza Υ de \mathbf{X} es un operador tal que:

$$\Upsilon : L_2([a, b])^p \rightarrow L_2([a, b])^p \quad (1.46)$$

$$\mathbf{f} \rightarrow \Upsilon(\mathbf{f}) = \int_a^b V(\cdot, t) \mathbf{f}(t) dt. \quad (1.47)$$

El operador de covarianza Υ es un operador integral con kernel V definido por:

$$V(s, t) = E [(\mathbf{X}(s) - \boldsymbol{\mu}(s)) \otimes (\mathbf{X}(t) - \boldsymbol{\mu}(t))], \quad s, t \in [a, b], \quad (1.48)$$

donde \otimes es el producto tensorial en \mathbb{R}^p . Así, para cualquier $s, t \in [a, b]$, $V(s, t)$ es una matriz de tamaño $p \times p$ con elementos [Jacques and Preda, 2014]:

$$V(s, t)[j, l] = Cov(X^j(s), X^l(t)). \quad (1.49)$$

1.2.2. Suavizado

Sea el proceso estocástico continuo $\mathbf{X} = \{\mathbf{X}(t)\}_{t \in [a, b]}$ con $\mathbf{X} = (X^1, X^2, \dots, X^p)^T \in \mathbb{R}^p$, para $p \geq 2$. Cada uno de los procesos X^i para $i = 1, 2, \dots, p$ es un dato funcional univariado representado por una curva simple. Una observación de \mathbf{X} se conoce con el

nombre de dato funcional multivariado, el cual está representado por un conjunto de p curvas que pertenecen a un espacio infinito-dimensional observadas en el intervalo compacto $I = [a, b]$. La dependencia entre estas p provee la estructura de \mathbf{X} [Jacques and Preda, 2014].

Teóricamente la observación del proceso \mathbf{X} está conformada por p funciones. Sin embargo, en la vida real se tiene que estas funciones no son observadas plenamente en el intervalo donde están definidas sino en puntos de este. Por consiguiente, es necesario ajustar curvas a este conjunto de puntos discretizados. Sea $\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_n$ una muestra independiente e idénticamente distribuida del proceso \mathbf{X} , el conjunto de datos observados son n vectores independientes de funciones $\mathbf{x}_i = (x_{i1}, x_{i2}, \dots, x_{ip})^T$ para $t \in I = [a, b]$, $i = 1, 2, \dots, n$.

Para esta tesis se supondrá que $\mathbf{X} \in L_2^p(I)$, donde $L_2^p(I) = L_2(I) \times L_2(I) \times \dots \times L_2(I)$ es el producto cruzado entre los espacios de Hilbert $L_2(I)$ de las p funciones cuadrado integrables en el intervalo compacto I , con el producto interno definido en 1.1. Esta función multivariada resulta cuadrado integrable dado que este producto de funciones univariadas es finito. En este trabajo se realizará el suavizamiento de las observaciones discretizadas empleando el método de suavizamiento spline [Ramsay and Silverman, 2005].

En total hay np funciones en la muestra. Estas se pueden representar por un número finito de bases de funciones ϕ_b . Para cada curva de la muestra aleatoria x_i^j , $j = 1, 2, \dots, p$ $i = 1, 2, \dots, n$, asumimos que se puede expresar como combinación lineal de bases de funciones $\{\phi_l^j\}_{j=1, B_l}$ como en la ecuación 1.50 [Jacques and Preda, 2014].

$$x_i^j(t) = \sum_{l=1}^{B_j} c_{ijl} \phi_l^j(t), \quad t \in I, \quad j = 1, 2, \dots, p \quad i = 1, 2, \dots, n. \quad (1.50)$$

Sea

$$\Phi(t) = \begin{bmatrix} \phi_{B_1}^1 & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \phi_{B_2}^2 & \dots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & \phi_{B_p}^p \end{bmatrix}, \quad (1.51)$$

la matriz cuyos elementos son las bases de las funciones y donde

$$\phi_{B_d}^j = (\phi_1^j(t), \dots, \phi_{B_d}^j(t)) \quad d = 1, \dots, p \quad j = 1, \dots, p. \quad (1.52)$$

Entonces se cumple que el vector de curvas \mathbf{x} se puede reescribir en términos de las bases de funciones como sigue:

$$\mathbf{x}_i(t) = \Phi(t) \mathbf{c}_i^T, \quad t \in I, \quad (1.53)$$

donde $\mathbf{c}_i = (c_{i11}, \dots, c_{i1B_1}, c_{i21}, \dots, c_{i2B_2}, \dots, c_{ip1}, \dots, c_{ipB_p})$ es el vector con los coeficientes de expansión de las bases.

1.2.3. Profundidad para datos funcionales multivariados

La profundidad estadística en datos funcionales es una medida de la centralidad de una observación (curva) respecto a un grupo de curvas o vectores de curvas, la cual

provee un ordenamiento del centro hacia afuera de las observaciones funcionales. Así, la curva con mayor profundidad es la curva más central y aquellas con menor profundidad se encuentran más alejadas, llegando incluso a ser datos funcionales atípicos. Como se describe en López-Pintado and Romo [2009], la curva más central se define como la curva mediana y a partir de esta estadística se pueden definir el concepto de percentiles en datos funcionales univariados y multivariados.

En datos funcionales multivariados, la profundidad se calcula empleando un conjunto de pesos definidos para cada uno de los p procesos que componen el dato multivariado, el cual es calculado de múltiples formas de acuerdo a las metodologías desarrolladas hasta el momento [Tarabelloni et al., 2014], [Claeskens et al., 2014], [Ieva and Paganoni, 2013]. Bajo esta definición, también se obtiene una estadística que permite realizar un ordenamiento de los datos funcionales multivariados, de acuerdo a su centralidad comparando uno a uno toda la muestra de vectores de curvas. A continuación, se definirá la profundidad de Claeskens para datos funcionales multivariados.

1.2.3.1. Profundidad funcional multivariada de Claeskens

Como se mencionó inicialmente, las profundidades para datos funcionales multivariados emplean un conjunto de pesos definidos para cada uno de los p procesos que componen el dato. En particular, Claeskens et al. [2014] consideraron la profundidad multivariada halfspace de Tukey asociada a unos pesos que se definirán a continuación.

Sea el proceso estocástico p -variado \mathbf{X} de funciones continuas definidas sobre el intervalo $I = [a, b]$, con función de distribución acumulativa $F_{\mathbf{X}}$. Para cada $t \in [a, b]$,

$$\mathbf{X}(t) = (X^1(t), X^2(t), \dots, X^p(t))^T, \quad (1.54)$$

es un vector aleatorio con función de distribución acumulativa $F_{\mathbf{X}}$.

Sea D una función de profundidad multivariada sobre \mathbb{R}^p y w una función de pesos que es definida sobre el intervalo $[a, b]$ y que integra 1. Si se toma un vector de curvas arbitrarias $\mathbf{X} \in C([a, b]^p)$ entonces la profundidad funcional multivariada (MFD) de \mathbf{X}^* se define en la ecuación 1.55.

$$MFD(\mathbf{X}^*, F_{\mathbf{X}}) = \int_a^b D(\mathbf{X}^*(t), F_{\mathbf{X}(t)})w(t)dt. \quad (1.55)$$

Un ejemplo de la función de pesos es la que toma los cambios locales en la cantidad de variabilidad en amplitud (variabilidad vertical), es decir

$$w(t) = w_{\alpha}(t, F_{\mathbf{X}(t)}) = \text{vol} \{D_{\alpha}(F_{\mathbf{X}(t)})\} / \int_a^b \text{vol} \{D_{\alpha}(F_{\mathbf{X}(u)})\} du, \quad (1.56)$$

el cual es proporcional al volumen de la región de profundidad en el tiempo t . Para definir la región de profundidad $D_{\alpha}(F_{\mathbf{X}(t)})$ se debe tener en cuenta las siguientes definiciones [Liu et al., 1999]:

Definición 1 . El conjunto $\{\mathbf{x} \in \mathbb{R}^p : D(\mathbf{x}; F_{\mathbf{X}(t)}) = h\}$ se conoce como el contorno de profundidad h .

Definición 2 . El conjunto $R(h) = \{\mathbf{x} \in \mathbb{R}^p : D(\mathbf{x}; F_{\mathbf{X}(t)}) > h\}$ se conoce como la región encerrada por el *contorno de profundidad* h .

Definición 3 . El conjunto $D_\alpha(F_{\mathbf{X}(t)}) = \bigcap_h \{R(h) : F_{\mathbf{X}(t)}(R(h)) \geq \alpha\}$ se conoce como la α -ésima región central. En otras palabras, $D_\alpha(F_{\mathbf{X}(t)})$ es la región más pequeña encerrada por contornos de profundidad para acumular una probabilidad α .

Un ejemplo de una función de profundidad multivariada D sobre \mathbb{R}^p es la profundidad halfspace de Tukey. Para un vector aleatorio $\mathbf{Y} \in \mathbb{R}^p$ con función de distribución acumulativa $F_{\mathbf{Y}}$ y una observación $\mathbf{y} \in \mathbb{R}^p$ de \mathbf{Y} , la profundidad halfspace de Tukey poblacional se define en la ecuación 1.57.

$$HD(\mathbf{y}, F_{\mathbf{Y}}) = \inf_{\mathbf{u} \in \mathbb{R}^p, \|\mathbf{u}\|=1} P(\mathbf{u}^T \mathbf{Y} \geq \mathbf{u}^T \mathbf{y}). \quad (1.57)$$

1.2.4. Medida atipicidad ajustada sobre datos funcionales multivariados

La atipicidad se define como una medida del grado de *atipicidad* de una observación respecto a la mediana de todo el conjunto de datos. En datos funcionales, una curva no atípica presenta valores de atipicidad que tienden a 0. A medida que este valor aumente, es decir que tienda a infinito, las curvas tenderán a alejarse de la distribución del conjunto de observaciones y se asociarán a datos atípicos. Esta medida, al igual que la profundidad, permite un ordenamiento de las curvas respecto a su centralidad. La relación entre una profundidad D y la atipicidad O , con rango en $[0, \infty)$, es inversa y se presenta a continuación [Hubert, 2008].

$$D = \frac{1}{1 + O}. \quad (1.58)$$

El objetivo de emplear esta medida, en datos funcionales, es identificar qué curvas presentan un comportamiento atípico respecto a las demás curvas, teniendo en cuenta que puede ser atípica, por ejemplo, por cambios de magnitud o de forma en las curvas.

En esta sección se inicia con una contextualización de la atipicidad en datos multivariados y luego una extensión a datos funcionales multivariados. Aunque existen muchas maneras de calcular la atipicidad de un dato respecto a un conjunto de datos [Brys et al., 2005], [Hubert, 2008], [Dai and Genton, 2019], aquí se presentarán las propuestas empleadas por Brys et al. [2005] y Hubert et al. [2017] referentes a atipicidad univariada para datos multivariados, atipicidad ajustada para datos multivariados y ajustada funcional para vectores de curvas.

1.2.4.1. Atipicidad univariada

Sea z una observación de una variable aleatoria **univariada** Z con distribución de probabilidad P_Z . Se define la atipicidad univariada AO_1 [Brys et al., 2005] como en la

ecuación (1.59).

$$AO_1(z, P_Z) = \begin{cases} \frac{z - \text{med}(Z)}{w_2(Z) - \text{med}(Z)} & \text{si } z > \text{med}(Z), \\ \frac{\text{med}(Z) - z}{\text{med}(Z) - w_1(Z)} & \text{si } z \leq \text{med}(Z), \end{cases} \quad (1.59)$$

donde $\text{med}(Z)$ es la mediana de Z , Q_1 y Q_3 son los cuartiles 1 y 3 de Z , IQR es el rango inter cuartílico de Z y $w_1(Z), w_2(Z)$ se definen como sigue:

$$w_1(Z) = Q_1(Z) - 1.5 \exp^{-4\text{MC}(Z)} \text{IQR}(Z),$$

$$w_2(Z) = Q_3(Z) + 1.5 \exp^{3\text{MC}(Z)} \text{IQR}(Z),$$

y $\text{MC}(Z)$ es una medida robusta de la asimetría (medcouple) propuesta en Brys et al. [2004] para un conjunto de observaciones z_1, z_2, \dots, z_n de la variable aleatoria Z , definida a continuación:

$$\text{MC}(z_1, z_2, \dots, z_n) = \text{med}_{i,j} \left(\frac{(z_j - \text{med}_k z_k) - (z_i - \text{med}_k z_k)}{z_j - z_i} \right), \quad (1.60)$$

y donde i, j deben cumplir

$$z_i \leq \text{med}_k z_k \leq z_j \quad z_i \neq z_j. \quad (1.61)$$

Cuando $\text{MC}(Z) < 0$ se reemplaza z por $-z$. El denominador de la ecuación 1.59 corresponde a los bigotes de boxplot ajustado multivariado, propuesto por Hubert and Vandervieren (2008) mencionados en Hubert et al. [2015]. Se puede observar que la medida AO_1 definida en la ecuación (1.59) tenderá a 0 cuando las observaciones se acerquen a la mediana, es decir, cuando las observaciones no presenten una dispersión alta. También tenderá a 0 cuando estos denominadores asociados a los bigotes del boxplot sean muy grandes, es decir, cuando las diferencias entre los bigotes y la mediana de la distribución sea grande implicando una variación alta de los datos. Por el contrario, las observaciones atípicas, es decir, con AO_1 muy grande, serán las que se encuentren lejos respecto a la mediana y para las cuales la distancia entre los bigotes y la mediana no sea alta, haciendo que la ecuación AO_1 presente valores grandes.

1.2.4.2. Atipicidad ajustada

Sea $\mathbf{x} \in \mathbb{R}^p$ un punto **multivariado** y $P_{\mathbf{Y}}$ la distribución de un vector aleatorio \mathbf{Y} en \mathbb{R}^p . Se define la *atipicidad ajustada* en la siguiente ecuación [Brys et al., 2005].

$$AO(\mathbf{x}; P_{\mathbf{Y}}) = \sup_{\|\mathbf{v}\|=1} AO_1(\mathbf{v}^T \mathbf{x}; P_{\mathbf{v}^T \mathbf{Y}}), \quad (1.62)$$

donde AO_1 es la atipicidad univariada descrita anteriormente.

Esta definición de atipicidad multivariada se interpreta como la máxima atipicidad univariada de las proyecciones unidimensionales de cada punto del conjunto. La medida AO se calcula a partir de AO_1 , en la cual una curva atípica se caracteriza por tener un valor lejano respecto a su mediana. En este caso se calcula la medida de forma análoga, aunque no se realizan cálculos sobre la nube de puntos, sí se realizan sobre la proyección

en una dimensión del conjunto de datos. Esto implicaría que en el caso multivariado, la medida AO también asociaría valores bajos a los vectores cuyas proyecciones no se alejen de la mediana, mientras que asocia valores altos a los vectores cuyas proyecciones se alejan de la mediana y los datos no presentan una dispersión tan alta.

1.2.4.3. Atipicidad ajustada funcional

Sea $\mathbf{Y} = \{\mathbf{Y}(t), t \in I\}$ un proceso estocástico p -variado continuo y sea $F_{\mathbf{Y}(t)}$ la distribución de \mathbf{Y} en el tiempo t . Se define la atipicidad ajustada funcional fAO [Hubert et al., 2017] de un vector de curvas \mathbf{X} sobre I respecto a \mathbf{Y} como en la ecuación (1.63).

$$\text{fAO}(\mathbf{X}; \mathbf{Y}) = \int_I \text{AO}(\mathbf{X}(t); F_{\mathbf{Y}(t)}) dt. \quad (1.63)$$

Se puede observar que el cálculo de la atipicidad para datos funcionales multivariados se realiza a partir de la definición de la misma en datos multivariados (ecuación (1.62)), integrando sobre el dominio en el que se definió el dato funcional multivariado. Esto implica que, para datos funcionales multivariados, la interpretación de esta estadística es análoga a la realizada en datos multivariados. Aquí, un vector de curvas con atipicidad alta, respecto a todo el conjunto de vectores de curvas, refleja que puede venir generado por otro proceso diferente, asociándolo a un dato atípico o en términos de control de calidad, un vector de curvas que se encuentra fuera de control.

1.2.5. Componentes principales funcionales multivariadas

Cuando se trabajan datos con muchas dimensiones, generalmente se opta por realizar una reducción de dimensión con la menor pérdida de información posible. Esto con el objetivo de evidenciar qué características se relacionan en el conjunto de datos y realizar un conjunto de transformaciones sobre las variables, que permitan trabajar en espacios de menor dimensión con nuevas variables independientes entre sí. En el análisis de datos funcionales esta tarea es muy relevante, dado que al trabajar con dimensión infinita esta reducción a un conjunto finito de observaciones facilita su visualización y optimiza el costo computacional.

1.2.5.1. Componentes principales para datos funcionales multivariados MFP-CA

Sea $\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_n$ una muestra independiente idénticamente distribuida de un dato funcional multivariado \mathbf{X} . La observación de $\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_n$ provee un conjunto de datos funcionales multivariados. Bajo la hipótesis de que éstas curvas son L_2 -continuas, es decir, cumplen la condición de la ecuación 1.44, el operador de covarianza Υ es un operador de Hilbert-Schmidt, esto significa que es compacto y auto adjunto, por consiguiente existe un conjunto de valores $\{\lambda_j\}_{j \geq 1}$ tal que [Jacques and Preda, 2014]:

$$- \sum_{j \geq 1} \lambda_j^2 < \infty$$

- $\{\lambda_j\}_{j \geq 1}$ son un conjunto de eigenvalores positivos asociados a una base ortonormal de funciones propias multivariadas $\{\mathbf{f}_j\}_{j \geq 1}$ para $\mathbf{f}_j = (f_j^1, f_j^2, \dots, f_j^p)$, llamadas factores principales o componentes principales y que son solución de

$$\Upsilon \mathbf{f}_j = \lambda_j \mathbf{f}_j, \quad (1.64)$$

con $\lambda_1 \geq \lambda_2 \geq \dots$ y $\int_a^b \sum_{l=1}^p f_j^l(t) f_{j'}^l(t) dt = 1$ si $j = j'$ y 0 en otro caso.

Los scores C_j de \mathbf{X} son variables aleatorias de media cero definidas como las proyecciones de \mathbf{X} sobre las funciones propias multivariadas de Υ [Jacques and Preda, 2014]:

$$C_j = \int_a^b \langle \mathbf{X}(t) - \boldsymbol{\mu}(t), \mathbf{f}_j(t) \rangle_{\mathbb{R}^p} dt = \int_a^b \sum_{l=1}^p (X^l(t) - \mu^l(t)) f_j^l(t) dt. \quad (1.65)$$

Similar al caso funcional univariado, los scores $\{C_j\}_{j \geq 1}$ son variables aleatorias con varianza λ_j con $j \geq 1$ [Jacques and Preda, 2014].

En el contexto multidimensional, la expansión de Karhunen-Loève [Jacques and Preda, 2014] queda definida de la siguiente manera:

$$\mathbf{X}(t) = \boldsymbol{\mu}(t) + \sum_{j \geq 1} C_j \mathbf{f}_j(t), \quad t \in [a, b]. \quad (1.66)$$

Para conjuntos de datos funcionales multivariados donde sus componentes tienen diferentes sistemas de unidades, se debe usar la versión normalizada de las componentes principales propuestas por Jacques and Preda [2014], las cuales se describen a continuación.

De forma similar al análisis de componentes clásico, la normalización se realiza introduciendo algunas métricas. Una de las formas para realizar esto en MFPCA es como en el análisis canónico, en el cual las componentes principales se definen como solución del siguiente eigen-problema [Jacques and Preda, 2014]:

$$\int_a^b \mathcal{P}_t(C_j) dt = \lambda_j C_j, \quad j \geq 1, \quad (1.67)$$

donde \mathcal{P}_t es el operador de proyección ortogonal asociado con \mathbf{X} , definido como

$$\mathcal{P}_t(C_j) = \langle \mathbf{X}(t), [V(t, t)]^{-1} \mathbb{E}[\mathbf{X}(t) C_j] \rangle_{\mathbb{R}^p}. \quad (1.68)$$

Combinando 1.67 y 1.68, se obtiene

$$C_j = \int_a^b \langle \mathbf{X}(t) - \boldsymbol{\mu}(t), \mathbf{f}_j(t) \rangle_{\mathbb{R}^p} dt, \quad (1.69)$$

donde \mathbf{f}_j es la solución del siguiente problema de eigenvectores

$$\int_a^b [V(s, s)]^{-1} [V(s, t)] \mathbf{f}_j(t) dt = \lambda \mathbf{f}(s). \quad (1.70)$$

Claramente, $[V(s, s)]^{-1}$ debe existir para cada $s \in [a, b]$. Bajo esta hipótesis, el factor principal del MFPCA normalizado son las eigenfunciones del operador integral con kernel $[V(s, s)]^{-1}[V(s, t)]$. La expansión de Karhunen-Loeve de \mathbf{X} será

$$\mathbf{X}(t) = \boldsymbol{\mu}(t) + \sum_{j=1}^{\infty} C_j [V(t, t)] \mathbf{f}_j(t), \quad t \in [a, b], \quad (1.71)$$

donde los score C_j , definidos en 1.69, tienen media cero y varianza λ_j .

1.2.5.2. Estimación de los componentes principales funcionales multivariados de \mathbf{X}

Sea \mathbf{X} un dato funcional multivariado, donde cada observación se compone de un vector finito dimensional cuyos elementos son p curvas infinito dimensionales, y sea $\mathbf{x}_1, \dots, \mathbf{x}_n$ un conjunto de n observaciones de \mathbf{X} . Se estiman las componentes principales funcionales como sigue [Happ and Greven, 2018]:

- Para cada elemento de $X^{(j)}$, $j = 1, \dots, p$ estime las componentes principales funcionales univariadas basado en las observaciones $x_1^{(j)}, \dots, x_n^{(j)}$. De esta reducción de dimensión se obtienen las funciones propias multivariadas estimadas $\hat{\phi}_m^{(j)}$ y los scores estimados $\hat{\xi}_{i,m}^{(j)}$ para $i = 1, \dots, n$ $m = 1, \dots, M_j$, donde M_j es el número de componentes principales seleccionadas para reducir la dimensión del proceso j .
- Definir la matriz $\Xi \in \mathbb{R}^{n \times M_+}$, $M_+ = M_1 + \dots + M_p$, donde cada fila $(\hat{\xi}_{i,1}^{(1)}, \dots, \hat{\xi}_{i,M_1}^{(1)}, \dots, \hat{\xi}_{i,1}^{(p)}, \dots, \hat{\xi}_{i,M_p}^{(p)})$ $i = 1, \dots, n$ contiene todos los scores estimados para cada una de las n observaciones del dato funcional multivariado. Luego se calcula la matriz $\hat{Z} = (n-1)^{-1} \Xi^T \Xi$ de dimensión $M_+ \times M_+$.
- Calcular los eigenvalores $\hat{\lambda}_m$ y los eigenvectores ortonormales \hat{c}_m para $m = 1, \dots, M_+$.
- Las funciones propias multivariadas estimadas están dadas por los elementos

$$\hat{f}_m^{(j)}(t) = \sum_{l=1}^{M_j} [\hat{c}_m]_l^{(j)} \hat{\phi}_l^{(j)}(t) \quad t \in I = [a, b] \quad m = 1, \dots, M_+, \quad (1.72)$$

y los scores multivariados estimados

$$\hat{C}_{i,m} = \sum_{j=1}^p \sum_{l=1}^{M_j} [\hat{c}_m]_l^{(j)} \hat{\xi}_{i,l}^{(j)} = \Xi_i \hat{c}_m. \quad (1.73)$$

Resulta muy importante definir un criterio para seleccionar el valor M_j con el objetivo de reducir la dimensión de cada proceso, el cual puede ser igual o diferente para cada $j = 1, 2, \dots, p$. Para esto es importante una revisión sobre los métodos explicados en análisis de componentes principales funcionales univariados.

1.2.6. Carta de control RSREWMA propuesta por Pan et al. [2019]

Esta carta fue desarrollada para detectar cambios en procesos para perfiles no lineales multivariados en fase II, empleando *regresión de vectores de soporte*. Esta regresión, conocida como *SVR*, es un algoritmo de aprendizaje estadístico supervisado para problemas de regresión. Aquí, las variables explicatorias son mapeadas sobre un espacio de características, y el modelo lineal descrito a continuación se construye sobre dicho espacio [Pan et al., 2019]

$$g(\mathbf{X}, w) = \mathbf{w}^T \phi(\mathbf{X}) + \mathbf{b}, \quad (1.74)$$

donde \mathbf{w} es un vector normal, $\phi()$ es una función de transformación no lineal y \mathbf{b} es el sesgo. La calidad de la estimación es medida por la función de pérdida $L(y_i, g(\mathbf{X}_i, \mathbf{w}))$, y el modelo *SVR* es formulado dentro de un problema de minimización como se define a continuación:

$$\min \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{i=1}^n (\xi_i + \xi_i^*), \quad (1.75)$$

sujeto a

$$\begin{cases} y_i - g(\mathbf{X}_i, \mathbf{w}) \leq \epsilon + \xi_i, \\ g(\mathbf{X}_i, \mathbf{w}) - y_i \leq \epsilon + \xi_i^*, \\ \xi_i, \xi_i^* \geq 0, \end{cases} \quad (1.76)$$

donde $\epsilon > 0$ es un umbral, la constante $C > 0$ es una penalización que puede ser vista como una forma de controlar el sobre-ajuste y ξ_i, ξ_i^* son variables de holgura. La función de pérdida viene dada por [Pan et al., 2019]

$$L(y_i, g(x_i, \mathbf{w})) = \begin{cases} 0 & \text{si } |y_i - g(x_i, \mathbf{w})| < \epsilon, \\ |y_i - g(x_i, \mathbf{w})| - \epsilon & \text{en otro caso} \end{cases} \quad (1.77)$$

La carta propuesta por Pan et al. [2019] es una carta EWMA (Exponentially Weighted Moving Average), combinada con el método del rango espacial y basada en la carta SREWMA, propuesta por Zou et. al. mencionados en Pan et al. [2019].

Para la carta propuesta por Pan et al. [2019], la cual denominan SREWMA no paramétrica revisada (RSREWMA), se evalúan las siguientes métricas, con el objetivo de comparar los perfiles de las muestras históricas en fase I con respecto a los perfiles observados en fase II, por medio del ajuste hecho con SVR:

1. SAD

$$w_{jk1} = \sum_{i=1}^{n_j} |y_{ijk} - \hat{y}_{ik}|. \quad (1.78)$$

2. MAD

$$w_{jk2} = \frac{1}{n_j} \sum_{i=1}^{n_j} |y_{ijk} - \hat{y}_{ik}|. \quad (1.79)$$

3. SSD

$$w_{jk3} = \sum_{i=1}^{n_j} (y_{ijk} - \hat{y}_{ik})^2. \quad (1.80)$$

donde $i = 1, 2, \dots, n_j$ es el tamaño de muestra, $k = 1, 2, \dots, p$ es el número de elementos que componen el vector de curvas, la respuesta y_{ijk} es el valor observado de la variable explicativa x_{ij} con la k -ésima característica de calidad en el perfil j y x_i es la variable explicativa correspondiente, tal que $i = 1, 2, \dots, n_j$ para cada $j = 1, 2, \dots$, y \hat{y}_{ik} es la estimación del perfil y_{ijk} empleando el método de la regresión de vectores de soporte descrita en la ecuación (1.74), es decir, $\hat{y}_{ik} = g(\mathbf{x}_i, \mathbf{w})$.

Se define el vector de métricas

$$\mathbf{w}_{jd} = (w_{j1d}, w_{j2d}, \dots, w_{jpd})^T; \quad j = 1, 2, \dots, \quad d = 1, 2, 3. \quad (1.81)$$

Sea $\{\mathbf{w}_{1d}^*, \mathbf{w}_{2d}^*, \dots, \mathbf{w}_{m_0d}^*\}$ el conjunto histórico de datos con tamaño m_0 , $d = 1, 2, 3$, los cuales son calculados desde los m_0 datos de los perfiles no lineales multivariados bajo control en fase I. Sea \mathbf{w}_{jd} , $j = 1, 2, \dots$, el vector de la d -ésima métrica para el correspondiente $j - 1$ -ésimo perfil no lineal multivariado en fase II. El rango espacial de \mathbf{w}_{jd} se puede calcular de la siguiente manera [Pan et al., 2019]

$$R_E(\widehat{\mathbf{M}}_{j-1} \mathbf{w}_{jd}) = \frac{1}{(m_0 + j - 1)} \sum_{q=-m_0+1}^{j-1} U(\widehat{\mathbf{M}}_{j-1}(\mathbf{w}_{jd} - \mathbf{w}_{qd})), \quad (1.82)$$

donde $\mathbf{w}_{(-m_0+j^*)d} = \mathbf{w}_{j^*d}^*$, $j^* = 1, 2, \dots, m_0$, $U(\cdot)$ es la función del signo espacial, y se define como

$$U(\mathbf{x}) \xrightarrow{iid} \begin{cases} \|\mathbf{x}\|^{-1} \mathbf{x}, & \mathbf{x} \neq 0 \\ 0, & \mathbf{x} = 0 \end{cases}, \quad \|\mathbf{x}\| = (\mathbf{x}^T \mathbf{x})^{1/2}, \quad (1.83)$$

$\mathbf{x}_i \in \mathbb{R}^p$ es la j -ésima observación de un vector p -dimensional, $\widehat{\mathbf{M}}_{j-1}$ es la raíz del inverso de la matriz triangular de Cholesky de $\widehat{\mathbf{S}}_{j-1}$ y $\widehat{\mathbf{S}}_{j-1}$ es la matriz de covarianza muestral de los datos históricos en el punto j [Pan et al., 2019]. Si la observación es reemplazada por el rango espacial $R_E(\widehat{\mathbf{M}}_{j-1} \mathbf{w}_{jd})$, entonces la estadística de control de la carta RSREWMA se escribe como

$$Q_j^{Re} = \frac{2 - \lambda}{\lambda} \mathbf{v}_j^T \{\text{cov}[R_F(\mathbf{M} \mathbf{w}_{jd})]\}^{-1} \mathbf{v}_j, \quad j = 1, 2, \dots, \quad (1.84)$$

donde

$$\text{cov}[R_F(\mathbf{M} \mathbf{w}_{jd})] = E[\|R_F(\mathbf{M} \mathbf{w}_{jd})\|^2] \mathbf{I}_p/p, \quad (1.85)$$

$$\mathbf{v}_j = (1 - \lambda) \mathbf{v}_{j-1} + \lambda R_E(\widehat{\mathbf{M}}_{j-1} \mathbf{w}_{jd}); \quad \mathbf{v}_0 = 0. \quad (1.86)$$

Para efectos computacionales, los autores definen $E[\|R_F(\mathbf{M} \mathbf{w}_{jd})\|^2]$ y $R_E(\widehat{\mathbf{M}}_{j-1} \mathbf{w}_{jd})$ como en las ecuaciones (1.87) y (1.88), respectivamente.

$$E[\|R_F(\mathbf{M} \mathbf{w}_{jd})\|^2] \approx \frac{\left[\sum_{q=-m_0+1}^0 \|\overline{R_E}(\widehat{\mathbf{M}}_0 \mathbf{w}_{qd})\|^2 \sum_{q=1}^{j-1} j-1 \|R_E(\widehat{\mathbf{M}}_{q-1} \mathbf{w}_{qd})\|^2 \right]}{(m_0 + j - 1)}, \quad (1.87)$$

$$\overline{R_E}(\widehat{\mathbf{M}}_{j-1} \mathbf{w}_{jd}) = \frac{1}{(m_0 + j - 1)} \sum_{q=-m_0+1}^{j-1} U(\widehat{\mathbf{M}}_{j-1}(\mathbf{w}_{jd} - \mathbf{w}_{qd})). \quad (1.88)$$

Si $Q_j^{Re} > L$, entonces la carta RSREWMA emite una alarma. En caso contrario, el vector \mathbf{w}_{jd} se incluye en la muestra histórica y se sigue monitoreando vectores de funciones

en fase II. Note que se necesita un valor del límite de control L para calcular el rendimiento de las cartas. Para hacer una implementación práctica Pan et al. [2019] emplean valores tabulados para combinaciones de λ y p , dado que estas cartas están propuestas en fase II.

Propuestas metodológicas

En este capítulo se presentan las propuestas para monitorear perfiles no lineales multivariados en fase II mediante un enfoque de datos funcionales multivariados. En la gráfica 2.1 se puede observar en los tiempos t_1, t_2, \dots, t_m , para $m \in \mathbb{Z}^+$, un conjunto de m perfiles no lineales multivariados. Por otro lado, en la gráfica 2.2 se ilustra una aproximación funcional de los perfiles mencionados, donde se evidencia que cada observación es un vector de funciones infinito dimensionales. En esta gráfica también se puede observar el objetivo del monitoreo de perfiles multivariados mediante datos funcionales, el cual consiste en identificar si una nueva observación de un dato funcional multivariado, suavizada y comparada con un conjunto histórico de datos funcionales multivariados, proviene del mismo proceso o no. Por ejemplo, en la figura 2.2 se puede observar en los tiempos t_1 y t_2 dos vectores de curvas provenientes del mismo proceso estocástico multivariado, mientras que en el tiempo t_m se tiene un vector de curvas generadas por otro proceso estocástico continuo multivariado. Se presentarán 3 enfoques de datos

- raw.png

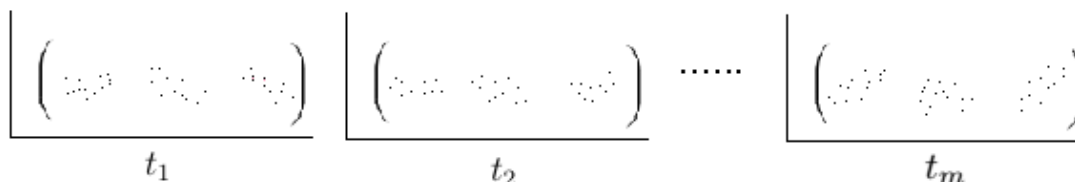


FIGURA 2.1. Observaciones de m perfiles no lineales multivariados

funcionales para monitorear perfiles no lineales, los cuales emplearán la profundidad propuesta por Claeskens [Claeskens et al., 2014], la medida atipicidad ajustada funcional [Hubert et al., 2017] y una propuesta de la semidistancia de Mahalanobis funcional basada en componentes principales funcionales multivariadas, que extiende la propuesta de Galeano et al. [2015].

Los enfoques propuestos en la literatura hasta ahora emplean estadísticas diferentes a las propuestas en esta Tesis para monitorear perfiles multivariados no lineales. Dentro

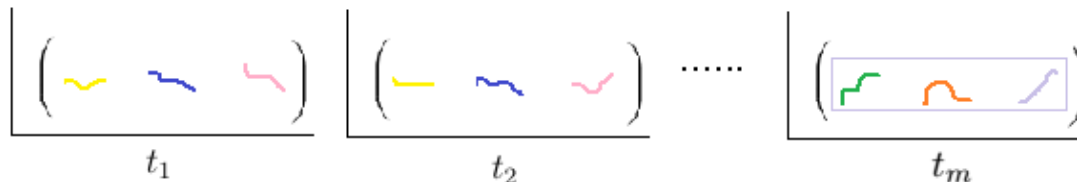


FIGURA 2.2. Aproximación funcional de un conjunto de m perfiles no lineales multivariados

de las metodologías existentes esta aquella que presenta restricciones respecto al comportamiento de los procesos, donde las funciones que conforman cada vector presentan comportamientos similares en fase I [Paynabar et al., 2016], o han realizado monitoreo mediante regresión no paramétrica usando *Support Vector Regression* [Pan et al., 2019]. También se han trabajado procesos debilmente correlacionados donde se emplearon componentes principales funcionales [Zhang et al., 2018], y el trabajo realizado por Ghosh et al. [2020] donde se emplean modelos de efectos mixtos ajustados mediante B-Splines para procesos gaussianos, pero no se han identificado trabajos con los enfoques propuestas en esta Tesis.

2.1. Modelo

Un perfil multivariado no lineal consiste en un conjunto de observaciones de un mismo individuo con una variable respuesta multivariada \mathbf{Y} y una o más variables explicativas \mathbf{X} . Para cada perfil $j = 1, 2, \dots$, se tienen las observaciones independientes $(\mathbf{X}_j, \mathbf{Y}_j)$, donde \mathbf{Y}_j es una matriz de tamaño $n_j \times p$ de variables respuesta y \mathbf{X}_j es un vector de tamaño $n_j \times 1$, n_j es el tamaño de muestra y puede tener valores diferentes para diferentes perfiles. Para cada valor de la(s) variable(s) explicativas(s), se tienen p valores de respuesta correspondientes. Cuando el proceso está bajo control, el modelo puede ser escrito como sigue [Pan et al., 2019]:

$$\mathbf{Y}_j = \mathbf{F}_j + \mathbf{E}_j; \quad j = 1, 2, \dots, \quad (2.1)$$

donde

$$\mathbf{E}_j = \begin{bmatrix} \mathbf{E}_{1j}^T \\ \mathbf{E}_{2j}^T \\ \vdots \\ \mathbf{E}_{n_j j}^T \end{bmatrix}, \quad \mathbf{E}_{ij}^T = (\epsilon_{ij1}, \dots, \epsilon_{ijp}), \quad (2.2)$$

o equivalentemente

$$\begin{bmatrix} y_{1j1} & \cdots & y_{1jp} \\ \vdots & \ddots & \vdots \\ y_{n_j j1} & \cdots & y_{n_j jp} \end{bmatrix}_{n_j \times p} = \begin{bmatrix} f_1(x_{1j}) & \cdots & f_p(x_{1j}) \\ \vdots & \ddots & \vdots \\ f_1(x_{n_j j}) & \cdots & f_p(x_{n_j j}) \end{bmatrix}_{n_j \times p} + \begin{bmatrix} \epsilon_{1j1} & \cdots & \epsilon_{1jp} \\ \vdots & \ddots & \vdots \\ \epsilon_{n_j j1} & \cdots & \epsilon_{n_j jp} \end{bmatrix}_{n_j \times p}, \quad (2.3)$$

donde $f_k(x_{ij})$ es una función con cierto grado de suavidad, la respuesta y_{ijk} es el valor observado de la variable explicativa x_i con la k -ésima característica de calidad en el perfil j , tal que $i = 1, 2, \dots, n_j$ para cada $j = 1, 2, \dots$. El número de características de calidad de interés es p , tal que $k = 1, 2, \dots, p$ y el término de error aleatorio \mathbf{E}_{ij}

generalmente asume alguna distribución multivariada [Pan et al., 2019]. El vector de funciones $\mathbf{f} = (f_1, \dots, f_p)^T$ será estimado usando suavizamiento spline, con el objetivo de ser aproximada por un estimador $\hat{\mathbf{f}}$ que será monitoreado empleando las estadísticas que se proponen en la siguiente sección. El vector de funciones \mathbf{f} será estimado a partir de las observaciones de cada perfil $(\mathbf{X}_j, \mathbf{Y}_j)$ dado que estas son observaciones discretas, y el objetivo es realizar el monitoreo de perfiles empleando técnicas de datos funcionales multivariados sobre vectores finitos de curvas infinito dimensionales.

Para esta Tesis se trabajará bajo el supuesto que $f_k \in L_2(I)$, $k = 1, 2, \dots, p$, donde $I = [a, b]$, $a < b$, $a, b \in \mathbb{R}$, es un intervalo compacto en los reales, de tal manera que $\mathbf{f} \in L_2^p(I)$, donde $L_2^p(I) = \underbrace{L_2(I) \times L_2(I) \times \dots \times L_2(I)}_{p \text{ veces}}$.

2.2. Enfoques propuestos

A continuación se presentan tres enfoques para monitorear perfiles no lineales multivariados empleando métodos de datos funcionales: Profundidad funcional multivariada, atipicidad ajustada funcional y componentes principales funcionales multivariadas. En cada uno de estos enfoques, se explica cómo se construye la carta de control incluyendo la estimación de los límites de control y cómo se monitorea el proceso en fase II.

2.2.1. Carta basada en profundidad funcional multivariada (Carta MFHD)

Sea $(\mathbf{x}_1, \mathbf{y}_1), \dots, (\mathbf{x}_m, \mathbf{y}_m)$ un conjunto de datos históricos bajo control que cumplen la condición de la ecuación 2.1, y sea $(\mathbf{x}_t, \mathbf{y}_t)$, para $t > m$, una nueva observación de un perfil multivariado que será monitoreado en fase II. Si el vector de curvas estimadas para el perfil $(\mathbf{x}_t, \mathbf{y}_t)$ es $\hat{\mathbf{f}}_t$, entonces se calculará la profundidad de $\hat{\mathbf{f}}_t$ con respecto a los demás vectores de curvas estimados en el conjunto histórico $\hat{\mathbf{f}}_1, \hat{\mathbf{f}}_2, \dots, \hat{\mathbf{f}}_m$.

Dado que la profundidad para datos funcionales multivariados permite ordenar del centro hacia afuera un conjunto de observaciones, de forma que el vector de curvas más profundo es el más central y el menos profundo es el más alejado respecto al centro, se calculará un límite que permita detectar si un vector de funciones se encuentra bajo control o no. Para esto se procede de la siguiente manera:

1. Establecer una tasa de falsas alarmas α , la cual es la probabilidad de que un vector de curvas sea identificado fuera de control cuando realmente está bajo control.
2. Calcular las profundidades de los vectores de curvas estimados en el conjunto histórico $\hat{\mathbf{f}}_1, \hat{\mathbf{f}}_2, \dots, \hat{\mathbf{f}}_m$.
3. Calcular el límite de control inferior, es decir el percentil q_α de las profundidades calculadas en el paso 2 mediante el método de bootstrap suavizado, es decir, $LCL = q_\alpha$, de la siguiente manera:

3.1 Sean $\hat{\mathbf{f}}_1, \hat{\mathbf{f}}_2, \dots, \hat{\mathbf{f}}_m$ un conjunto de m funciones estimadas a partir de las observaciones históricas de un conjunto de vectores de curvas y T la estadística de interés que se calculará sobre la muestra, la cual en este caso es el percentil q_α de las profundidades calculadas para cada vector de funciones respecto a todos los demás en la muestra.

3.2 Dado el conjunto de vectores de curvas suavizadas $(\hat{\mathbf{f}}_1, \hat{\mathbf{f}}_2, \dots, \hat{\mathbf{f}}_m)$, tomar un conjunto de B muestras con reemplazamiento (bootstrap) de estas. Si se denota una muestra genérica de estas como $\hat{\mathbf{f}}_1^*, \hat{\mathbf{f}}_2^*, \dots, \hat{\mathbf{f}}_m^*$, entonces se aplica un ruido a cada una de las p curvas del vector de la muestra bootstrap, para evitar que aparezcan medidas repetidas en las muestras artificiales, de la siguiente manera:

$$\hat{\mathbf{f}}_{ij}^b = \hat{\mathbf{f}}_{ij}^* + \mathbf{Z}_j, \quad i \in (1, 2, \dots, m), \quad j \in (1, 2, \dots, p), \quad (2.4)$$

donde \mathbf{Z}_j es un vector de observaciones provenientes de una variable aleatoria multivariada con distribución normal de media $\mathbf{0}$ y matriz de varianzas y covarianzas $\gamma \Sigma_x$, Σ_x es la matriz (de tamaño $p \times p$) de covarianza de $(\hat{\mathbf{f}}_1(t), \hat{\mathbf{f}}_2(t), \dots, \hat{\mathbf{f}}_m(t))$ y γ es el parámetro de suavizado, el cual Cuevas et al. [2006] recomiendan sea $\gamma = 0.05$.

En este caso, se define

$$\hat{\mathbf{f}}_i^b = (\hat{\mathbf{f}}_{i1}^b, \hat{\mathbf{f}}_{i2}^b, \dots, \hat{\mathbf{f}}_{ip}^b), \quad i \in (1, 2, \dots, m).$$

3.3 De acuerdo a la ecuación 1.55, se calculará la profundidad multivariada funcional de cada vector de curvas obtenido vía bootstrap suavizado, con respecto a todos los demás vectores de funciones bootstrap en la muestra y, posteriormente, el percentil α de estas profundidades:

$$T^b = q_\alpha \left(MFHD_1(\hat{\mathbf{f}}_1^b, F_{\hat{\mathbf{f}}_1^b, \dots, \hat{\mathbf{f}}_m^b}), MFHD_2(\hat{\mathbf{f}}_2^b, F_{\hat{\mathbf{f}}_1^b, \dots, \hat{\mathbf{f}}_m^b}), \dots, MFHD_m(\hat{\mathbf{f}}_m^b, F_{\hat{\mathbf{f}}_1^b, \dots, \hat{\mathbf{f}}_m^b}) \right),$$

empleando la b -ésima muestra bootstrap.

3.4 Una vez calculada la estadística de interés T^b , se repite el proceso B veces tomando finalmente el promedio como el estimador del parámetro de interés $\bar{T} = \sum_{b=1}^B \frac{T^b}{B}$. En este trabajo se emplean valores $B = 500$ repeticiones del método, que resultaron suficientes para encontrar estimaciones robustas.

En este caso se calcula un límite inferior de control (LCL) dado que entre más se aproxime a cero la profundidad de un vector de curvas, esto indica que se aleja más de los vectores de curvas centrales. No se calcula un límite superior pues, entre más grande sea el valor de la profundidad, más central será la observación, lo cual la hace menos probable a estar fuera de control.

Para realizar monitoreo de perfiles no lineales multivariados en fase II, se calcula la profundidad del nuevo vector de curvas $\hat{\mathbf{f}}_t$ con respecto al conjunto histórico de vectores de curvas estimados, usando la profundidad funcional multivariada *MFHD* propuesta por Claeskens et al. [2014], ver ecuación 1.55. La carta de control se construye graficando para cada t el valor de la estadística $MFHD(\hat{\mathbf{f}}_t, F_{\hat{\mathbf{f}}_1, \dots, \hat{\mathbf{f}}_m})$. Si la profundidad de este vector de curvas es mayor al LCL, entonces el proceso caracterizado por $\hat{\mathbf{f}}_t$ se encuentra bajo control, en caso contrario se encuentra fuera de control como se observa en la gráfica 2.3, donde se encerraron en círculos los dos puntos de profundidad funcional multivariada de vectores de curvas fuera de control.

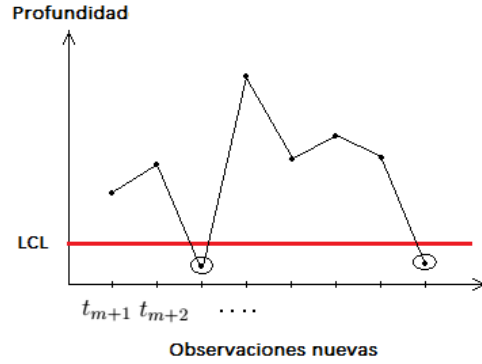


FIGURA 2.3. Carta de control bajo el enfoque de la profundidad multivariada MFHD para monitorear perfiles no lineales multivariados

En esta Tesis se calculará la profundidad de un vector de funciones $\hat{\mathbf{f}}_t$ con respecto al conjunto histórico $\hat{\mathbf{f}}_1, \dots, \hat{\mathbf{f}}_m$ usando la profundidad funcional multivariada propuesta por Claeskens et al. [2014]. Usando la ecuación 1.55, esta profundidad está dada por:

$$MFHD_t(\hat{\mathbf{f}}_t, F_{\hat{\mathbf{f}}_1, \dots, \hat{\mathbf{f}}_m}) = \int_I D(\hat{\mathbf{f}}_t(s), F_{\hat{\mathbf{f}}_1(s), \dots, \hat{\mathbf{f}}_m(s)}) w(s) ds, \quad (2.5)$$

donde $w(s)$ representa una función de pesos descrita en la ecuación 1.56, $s \in I$ y

$$D(\hat{\mathbf{f}}_t(s), F_{\hat{\mathbf{f}}_1(s), \dots, \hat{\mathbf{f}}_m(s)}) = \inf_{\mathbf{u} \in \mathbb{R}^p, \|\mathbf{u}\|=1} P(\mathbf{u}^T F_{\hat{\mathbf{f}}_1(s), \dots, \hat{\mathbf{f}}_m(s)} \geq \mathbf{u}^T \hat{\mathbf{f}}_t(s)). \quad (2.6)$$

La carta de control construida bajo ese enfoque se denomina carta MFHD.

2.2.2. Carta basada en la medida atipicidad ajustada funcional (Carta fAO)

Mediante el enfoque de la atipicidad ajustada funcional, el objetivo será estimar el límite de control superior (UCL) en la fase I mediante la técnica de bootstrap suavizado. En este caso se estima el límite superior, dado que un vector de curvas con un valor alto de atipicidad ajustada funcional indica que se aleja del comportamiento central del conjunto histórico y aumenta sus probabilidades de estar fuera de control. En la figura 2.4 se puede observar que, en fase II, uno de los vectores de curvas presentó una atipicidad ajustada funcional muy alta (encerrado en un círculo) respecto al conjunto histórico de datos y se

marcó como fuera de control. El siguiente procedimiento ilustra cómo se determina el UCL:

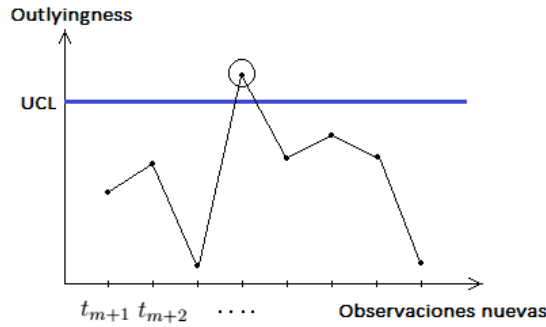


FIGURA 2.4. Carta de control bajo el enfoque de atipicidad ajustada funcional para monitorear perfiles no lineales multivariados

1. Establecer una tasa de falsas alarmas α , la cual es la probabilidad de que un vector de curvas sea identificado fuera de control cuando realmente está bajo control.
2. Calcular la atipicidad ajustada funcional de los vectores de curvas estimados en el conjunto histórico $\hat{\mathbf{f}}_1, \hat{\mathbf{f}}_2, \dots, \hat{\mathbf{f}}_m$.
3. Calcular el percentil $(1 - \alpha)$ de las atipicidades ajustadas funcionales calculadas en el paso 2 mediante el método de bootstrap suavizado, como se describe a continuación.
 - 3.1 Sean $\hat{\mathbf{f}}_1, \hat{\mathbf{f}}_2, \dots, \hat{\mathbf{f}}_m$ un conjunto de m funciones estimadas a partir de las observaciones históricas de un conjunto de vectores de curvas y T la estadística de interés que se calculará sobre la muestra, la cual en este caso es el percentil $q_{1-\alpha}$ de las atipicidades ajustadas funcionales calculadas para cada vector de funciones respecto a todos los demás en la muestra.
 - 3.2 Dado el conjunto de vectores de curvas suavizadas $(\hat{\mathbf{f}}_1, \hat{\mathbf{f}}_2, \dots, \hat{\mathbf{f}}_m)$, tomar un conjunto de B muestras con reemplazamiento (bootstrap) de estas. Si se denota una muestra genérica de estas como $\hat{\mathbf{f}}_1^*, \hat{\mathbf{f}}_2^*, \dots, \hat{\mathbf{f}}_m^*$, entonces se aplica un ruido a cada una de las p curvas del vector de la muestra bootstrap, para evitar que aparezcan medidas repetidas en las muestras artificiales, de la siguiente manera:

$$\hat{\mathbf{f}}_{ij}^b = \hat{\mathbf{f}}_{ij}^* + \mathbf{Z}_j, \quad i \in (1, 2, \dots, m), \quad j \in (1, 2, \dots, p), \quad (2.7)$$

donde \mathbf{Z}_j es un vector de observaciones provenientes de una variable aleatoria multivariada con distribución normal de media $\mathbf{0}$ y matriz de varianzas y covarianzas $\gamma \Sigma_x$, Σ_x es la matriz (de tamaño $p \times p$) de covarianza de $(\hat{\mathbf{f}}_1(t), \hat{\mathbf{f}}_2(t), \dots, \hat{\mathbf{f}}_m(t))$ y γ es el parámetro de suavizado, el cual Cuevas et al. [2006] recomiendan sea $\gamma = 0.05$.

En este caso, se define

$$\hat{\mathbf{f}}_i^b = \left(\hat{\mathbf{f}}_{i1}^b, \hat{\mathbf{f}}_{i2}^b, \dots, \hat{\mathbf{f}}_{ip}^b \right), \quad i \in (1, 2, \dots, m).$$

- 3.3 De acuerdo a la ecuación 1.63, se calculará la atipicidad ajustada funcional multivariada de cada vector de curvas con respecto a todos los demás en la muestra y, posteriormente, el percentil $(1 - \alpha)$ de estas medidas:

$$T^b = q_{1-\alpha} \left(\text{fAO}_1(\hat{\mathbf{f}}_1^b; \hat{\mathbf{f}}_1^b, \dots, \hat{\mathbf{f}}_m^b), \text{fAO}_2(\hat{\mathbf{f}}_2^b; \hat{\mathbf{f}}_1^b, \dots, \hat{\mathbf{f}}_m^b), \dots, \text{fAO}_m(\hat{\mathbf{f}}_m^b; \hat{\mathbf{f}}_1^b, \dots, \hat{\mathbf{f}}_m^b) \right),$$

empleando la b -ésima muestra bootstrap.

- 3.4 Una vez calculada la estadística de interés T^b se repite el proceso B veces tomando finalmente el promedio como el estimador del parámetro de interés $\bar{T} = \sum_{b=1}^B \frac{T^b}{B}$. En este trabajo se emplean valores $B = 500$ repeticiones del método.

En este caso, se calcula un límite superior de control (UCL) dado que, entre más grande sea el valor de la atipicidad ajustada funcional de un vector de curvas, esto indica que se aleja más de los vectores de curvas centrales. No es necesario encontrar un límite inferior de control dado que entre más tienda a 0 la medida será porque la observación presenta muy poca atipicidad respecto al conjunto histórico de datos.

4. Finalmente, se calcula la atipicidad ajustada funcional del nuevo vector de curvas $\hat{\mathbf{f}}_t$ respecto al conjunto histórico de vectores de curvas estimados de acuerdo a la ecuación 1.63:

$$\text{fAO}_t \left(\hat{\mathbf{f}}_t; \hat{\mathbf{f}}_1, \dots, \hat{\mathbf{f}}_m \right) = \int_I \text{AO} \left(\hat{\mathbf{f}}_t(s); F_{\hat{\mathbf{f}}_1(s), \dots, \hat{\mathbf{f}}_m(s)} \right) ds, \quad (2.8)$$

donde AO se define como en la ecuación 1.62 y $I = [a, b]$, $a < b$, $a, b \in \mathbb{R}$. Finalmente, se compara el valor obtenido en la ecuación 2.8 con el UCL calculado en el paso 3: Si la atipicidad ajustada funcional de este vector de funciones es menor al límite entonces la curva se encuentra bajo control, en caso contrario se encuentra fuera de control como se observa en la gráfica 2.4, donde se encerró en un círculo el punto de la estadística del vector de curvas fuera de control.

La carta de control construida bajo ese enfoque se denomina carta fAO.

2.2.3. Carta basada en la semidistancia de Mahalanobis funcional datos funcionales multivariados (Carta \mathbf{T}_{MF}^2)

De la misma forma que en la figura 2.4, aquí se calculará un UCL dado que una mayor semidistancia implica que el vector de funciones se aleja del conjunto histórico bajo control

y para cartas que emplean semidistancia T^2 automáticamente se define el límite de control inferior $LCL = 0$ [Montgomery, 2007] cómo se ve en la gráfica 2.5. Se propone una carta de control basada en la semidistancia de Mahalanobis funcional definida en la ecuación (1.40). Sin embargo, dado que esta medida fue propuesta para datos funcionales univariados [Galeano et al., 2015], se ajustará de forma análoga para datos funcionales multivariados con base en las componentes principales funcionales multivariadas propuestas por Happ and Greven [2018]. La metodología para construir la carta se describe a continuación.

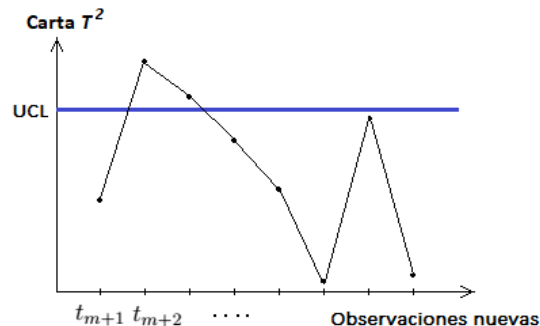


FIGURA 2.5. Carta de control bajo el enfoque de semidistancia T_{MF}^2 para monitorear perfiles no lineales multivariados

1. Se empieza en la fase I empleando un conjunto histórico de vectores de funciones bajo control estimadas $\hat{\mathbf{f}}_1, \hat{\mathbf{f}}_2, \dots, \hat{\mathbf{f}}_m$, es decir, que sean realizaciones del mismo proceso estocástico multivariado. Luego, se calculan las componentes principales multivariadas funcionales bajo el enfoque propuesto por Happ and Greven [2018]. De este análisis se obtiene un conjunto $\lambda_1 > \lambda_2 > \dots > \lambda_K > 0$ de eigenvalores y un conjunto de funciones propias multivariadas $\mathbf{g}_j = (g_j^1, g_j^2, \dots, g_j^p)$, $j = 1, \dots, K$, para un valor K que retenga un porcentaje alto de la varianza de los datos funcionales multivariados.

Primero se estima el vector de curvas promedio histórico bajo control del conjunto de m vectores de curvas de la siguiente manera:

$$\hat{\boldsymbol{\mu}}_{\hat{\mathbf{f}}}(s) = \frac{\sum_{j=1}^m \hat{\mathbf{f}}_j(s)}{m}, \quad s \in I, \quad (2.9)$$

donde I es un intervalo compacto en los reales.

El conjunto de scores $C_j : j = 1, \dots, K$, se calculan como se indicó en la ecuación 1.65, dado que la propuesta realizada por Jacques and Preda [2014] es un caso particular de la propuesta por Happ and Greven [2018] y es válido emplear la siguiente ecuación:

$$C_{j,h} = \int_I \langle \hat{\mathbf{f}}_h - \hat{\boldsymbol{\mu}}(s), \mathbf{g}_j(s) \rangle_{\mathbb{R}^p} ds = \int_I \sum_{l=1}^p (f_h^l(s) - \hat{\boldsymbol{\mu}}^l(s)) g_j^l(s) ds. \quad (2.10)$$

El valor de K fue determinado de acuerdo al criterio del porcentaje de varianza acumulada [Ramsay and Silverman, 2005] y el método del *scree plot*, descrito en la

sección de análisis de componentes principales funcionales.

2. Una vez calculados los scores, se procede a calcular la semidistancia de Mahalanobis de cada dato funcional multivariado respecto al vector de curvas promedio histórico estimado, $(\hat{\mathbf{f}}_h; \hat{\mathbf{f}}_1, \dots, \hat{\mathbf{f}}_m)$ [Galeano et al., 2015]:

$$T_{MF,h}^2 = \sum_{k=1}^K \frac{C_{k,h}^2}{\lambda_k}, \quad (2.11)$$

donde $C_{k,h}$ está definido por:

$$C_{k,h} = \int_I \langle \hat{\mathbf{f}}_h - \hat{\boldsymbol{\mu}}_{\hat{\mathbf{f}}}(s), \mathbf{g}_k(s) \rangle_{\mathbb{R}^p} ds = \int_I \sum_{l=1}^p (f_h^l(s) - \hat{\mu}_{\hat{\mathbf{f}}}^l(s)) g_k^l(s) ds. \quad (2.12)$$

La estadística T_{MF}^2 se calcula como la semidistancia de Mahalanobis clásica entre los scores de los datos funcionales multivariados, dado que la media de estos es 0 y la varianza de cada score $\mathbf{C}_{j,t}$ para $j = 1, 2, \dots, K$ es el eigenvalor λ_j para cada vector de curvas observado del proceso (ver figura 2.2). Si el número de procesos estocásticos es uno, $p = 1$, entonces el vector de funciones \mathbf{f}_h se notará en f_h , tendrá una sola función y la estadística se expresará como $T_{F,h}^2$, como se define a continuación:

$$T_{F,h}^2 = \sum_{k=1}^K \frac{c_{k,h}^2}{\lambda_k}, \quad (2.13)$$

donde

$$c_{k,h} = \int_I \langle f_h - \hat{\mu}_{\hat{f}}(s), g_k(s) \rangle_{\mathbb{R}} ds = \int_I (f_h(s) - \hat{\mu}_{\hat{f}}(s)) g_k(s) ds. \quad (2.14)$$

son los scores calculados bajo el análisis de componentes principales funcionales univariadas y $\hat{\mu}_{\hat{f}}$ es la media histórica, calculada de forma análoga a la ecuación 2.9.

En este caso, cuando $p = 1$ y el intervalo I es compacto en \mathbb{R} , las componentes principales funcionales multivariadas propuestas por Happ and Greven [2018], descritas en el capítulo anterior, son equivalentes a las componentes principales funcionales univariadas empleadas por Galeano et al. [2015] descritas en la sección 1.1.5.

3. A continuación, el UCL se calculará con el percentil $(1 - \alpha)$ de la distribución de T_{MF}^2 mediante la técnica de bootstrap suavizado descrita a continuación.

- 3.1 Sean $\hat{\mathbf{f}}_1, \hat{\mathbf{f}}_2, \dots, \hat{\mathbf{f}}_m$ el conjunto de m funciones estimadas a partir de las observaciones históricas de un conjunto de vectores de curvas y T la estadística de interés que se calculará sobre la muestra, la cual en este caso es el percentil $q_{1-\alpha}$ de las distancias T^2 calculada para cada vector de funciones respecto a todos los demás en la muestra.

- 3.2 Dado el conjunto de vectores de curvas suavizadas $(\hat{\mathbf{f}}_1, \hat{\mathbf{f}}_2, \dots, \hat{\mathbf{f}}_m)$, tomar un conjunto de B muestras con reemplazamiento (bootstrap) de estas. Si se denota

una muestra genérica de estas como $\hat{\mathbf{f}}_1^*, \hat{\mathbf{f}}_2^*, \dots, \hat{\mathbf{f}}_m^*$, entonces se aplica un ruido a cada una de las p curvas del vector de la muestra bootstrap, para evitar que aparezcan medidas repetidas en las muestras artificiales, de la siguiente manera:

$$\hat{\mathbf{f}}_{ij}^b = \hat{\mathbf{f}}_{ij}^* + \mathbf{Z}_j, \quad i \in (1, 2, \dots, m), \quad j \in (1, 2, \dots, p), \quad (2.15)$$

donde \mathbf{Z}_j es un vector de observaciones provenientes de una variable aleatoria multivariada con distribución normal de media $\mathbf{0}$ y matriz de varianzas y covarianzas $\gamma \Sigma_x$, Σ_x es la matriz (de tamaño $p \times p$) de covarianza de $(\hat{\mathbf{f}}_1(t), \hat{\mathbf{f}}_2(t), \dots, \hat{\mathbf{f}}_m(t))$ y γ es el parámetro de suavizado, el cual Cuevas et al. [2006] recomiendan sea $\gamma = 0.05$.

En este caso, se define

$$\hat{\mathbf{f}}_i^b = (\hat{\mathbf{f}}_{i1}^b, \hat{\mathbf{f}}_{i2}^b, \dots, \hat{\mathbf{f}}_{ip}^b), \quad i \in (1, 2, \dots, m).$$

- 3.3 De acuerdo a la ecuación 2.11, se calculará la semidistancia T^2 multivariada funcional de cada vector de curvas $\hat{\mathbf{f}}_i^b$, $i = 1, \dots, m$ respecto a todos los demás en la muestra y, posteriormente, el percentil $(1 - \alpha)$ de estas distancias:

$$T^b = q_{1-\alpha}(T_{MF,1}^2, T_{MF,2}^2, \dots, T_{MF,m}^2),$$

empleando la b -ésima muestra bootstrap.

- 3.4 Una vez calculada la estadística de interés T^b se repite el proceso B veces tomando finalmente el promedio como el estimador del parámetro de interés $\bar{T} = \sum_{b=1}^B \frac{T^b}{B}$. En este trabajo se emplean valores $B = 500$ replicaciones del método.

Como resultado asintótico de esta distribución, en Horváth and Kokoszka [2012] se observa que, cuando se trabaja con procesos gaussianos, la estadística T_{MF}^2 converge a una distribución χ^2 con K grados de libertad. Si el proceso no es gaussiano, se calcula el límite de control superior (UCL) en fase I, por medio de la técnica de bootstrap suavizado, con un conjunto de datos que se encuentran bajo control (fase I) y que servirán de base en fase II para detectar vectores de funciones fuera de control. En esta Tesis también compararemos el resultado del límite obtenido por bootstrap suavizado con el límite teórico obtenido de la distribución de la estadística T_{MF}^2 , para evaluar las propiedades de la estadística empleada cuando los procesos se generen a partir de procesos gaussianos, dado que es la condición requerida para la convergencia en distribución.

4. Finalmente, se calcula la semidistancia T_{MF}^2 de un nuevo vector de curvas $\hat{\mathbf{f}}_t$ respecto al conjunto histórico de vectores de curvas estimados, la media histórica y las funciones propias multivariadas históricas estimadas bajo control, de acuerdo a la ecuación 2.11:

$$T_{MF,t}^2 = \sum_{k=1}^K \frac{C_{k,t}^2}{\lambda_k}, \quad (2.16)$$

donde $C_{k,t}^2$ está definido por:

$$C_{k,t} = \int_I \langle \hat{\mathbf{f}}_t(s) - \hat{\boldsymbol{\mu}}_{\hat{\mathbf{f}}}(s), \mathbf{g}_k(s) \rangle_{\mathbb{R}^p} ds = \int_I \sum_{l=1}^p (f_t^l(s) - \hat{\mu}_{\hat{\mathbf{f}}}^l(s)) g_k^l(s) ds. \quad (2.17)$$

La carta de control construida bajo ese enfoque se denomina carta T_{MF}^2 .

Simulaciones

En este capítulo se presentan las simulaciones de las propuestas metodológicas explicadas en el capítulo anterior. Se inicia con la simulación de datos funcionales univariados para evaluarlos respecto a una versión multivariada empleando sus derivadas, con el objetivo de identificar si al incluir mayor información del proceso se puede realizar cartas de control más potentes. Luego se simulan datos funcionales multivariados para implementar las propuestas desarrolladas en la Tesis y evaluar cual presenta los mejores resultados.

Las simulaciones se realizarán de la siguiente manera:

- Inicialmente, se simulan conjuntos de datos funcionales sin contaminación para determinar los límites de control UCL y LCL bajo la metodología bootstrap suavizado.
- Seguido a esto, se simulan conjuntos de datos funcionales contaminados (en magnitud o forma). Finalmente, se compara el desempeño de las cartas propuestas en esta Tesis contra las propuestas por Sheu et al. [2013] y Pan et al. [2019].

3.1. Evaluación del rendimiento de la carta de control

Generalmente, el desempeño de una carta de control se evalúa usando la longitud promedio de corrida (Average Run Length, ARL por su sigla en inglés), el cual se define como el número promedio de muestras requeridas para detectar un cambio en el proceso. Para cartas con longitudes de corrida geométricas, cuando el proceso está bajo control el ARL es denotado como $ARL_0 = \frac{1}{\alpha}$, donde α representa la tasa de falsas alarmas en fase I. Cuando el proceso está fuera de control, el ARL es denotado como $ARL_1 = \frac{1}{1-\beta}$, donde β representa la probabilidad de no detectar cambios en el proceso [Montgomery, 2007].

En esta tesis se emplearán tanto la potencia como el ARL para medir el rendimiento de las cartas de control. La potencia se define como $p = 1 - \beta$, donde β es definida igual

que en el párrafo anterior, dado que para cartas tales como la F no es sencillo encontrar la potencia exacta de los límites de control [Sheu et al., 2013]. Ya que el ARL depende de un cálculo preciso de la potencia, dado que esta se encuentra en el denominador, una aproximación muy variable puede afectar su valor ya sea subestimado o sobreestimado.

3.2. Comparación cartas univariadas vs multivariadas

Una ventaja de considerar los perfiles como objetos funcionales es que estos pueden visualizar ciertos cambios, en especial, cuando son de forma. Además, ellos tienen información implícita que se exhibe derivando las funciones, que podría permitir identificar procesos fuera de control de forma más efectiva.

Sea X un dato funcional univariado que se ajusta a un conjunto de observaciones en puntos discretos de un intervalo $I = [a, b]$, $a < b$, $a, b \in \mathbb{R}$. Ahora se considera el proceso multivariado $\mathbf{X} = (X, X', X'')$, donde X' y X'' representan la primera y segunda derivada de X , respectivamente. De esta manera se transforma un dato funcional univariado en uno multivariado. Esta transformación enriquece la información del proceso original con la que aportan las primeras derivadas, de forma similar a lo que sucede en el cálculo diferencial.

En esta sección se presenta la simulación de un dato funcional univariado, al cual se le calculan la primera y segunda derivada, con el objetivo de comparar las cartas F , T_F^2 , T_{MF}^2 , fAO y MFHD.

Dado que las estadísticas T^2 funcionales, tanto univariada como multivariada, requieren una reducción en componentes principales funcionales (univariadas y multivariadas), se debe tener precaución a la hora de seleccionar el número de componentes a retener. Los resultados de los estudios realizados en Horváth and Kokoszka [2012] evidencian la importancia de seleccionar valores no tan altos dado que los eigenvalores tienden a 0 y al estar en un denominador aumentarán significativamente el valor de las estadísticas en la ecuación (2.11).

El modelo empleado para comparar las cartas, generando los conjuntos de perfiles univariados históricos es el propuesto en Galeano et al. [2015], definido en la siguiente ecuación.

$$\chi(t) = W(t) + \mu_\chi(t), \quad t \in [0, 1], \quad (3.1)$$

donde se define la función media del proceso $\mu_\chi(t) = 20t^{1.1}(1-t)$ y $W(t)$ es un movimiento browniano, cuyos valores propios están definidos en la ecuación (3.2) y las funciones propias en la ecuación (3.3) [Kokoszka and Reimherr, 2017].

$$\lambda_k = \frac{1}{(\pi(k - 0.5))^2}, \quad k = 1, 2, \dots, \quad (3.2)$$

$$\psi_k(t) = \sqrt{2} \sin((k - 0.5)\pi t), \quad k = 1, 2, \dots, \quad (3.3)$$

De acuerdo al modelo definido en la ecuación (3.1), se simulan las observaciones discretizadas y_i , definidas como:

$$y_i(t) = \chi_i(t) + \epsilon_i(t), \quad t \in [0, 1], \quad i = 1, 2, \dots, m, \quad (3.4)$$

donde m el número de observaciones históricas del dato funcional univariado y $\epsilon_i \sim N(0, 0.1)$, es decir, el error tiene una distribución normal con media 0 y varianza 0.01 (desviación $\sigma = 0.1$).

Para la simulación del proceso se toma $k = 1, 2, \dots, 10$ [Galeano et al., 2015]. En la Figura 3.1 se puede observar un conjunto de perfiles simulados bajo estas condiciones. Una vez se simulan los perfiles y se suavizan, se procede a generar el dato funcional

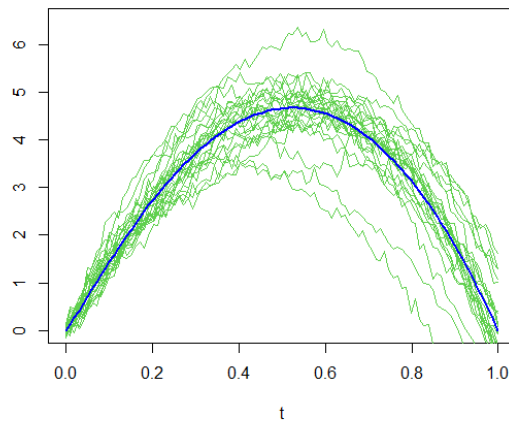


FIGURA 3.1. Muestra de observaciones discretizadas bajo el esquema de la ecuación (3.4) y en azul la función media μ_χ

multivariado calculando la primera y la segunda derivada, como se observa en la Figura 3.2. Se generarán 500 conjuntos de $m = 1000$ curvas en fase I para determinar los límites

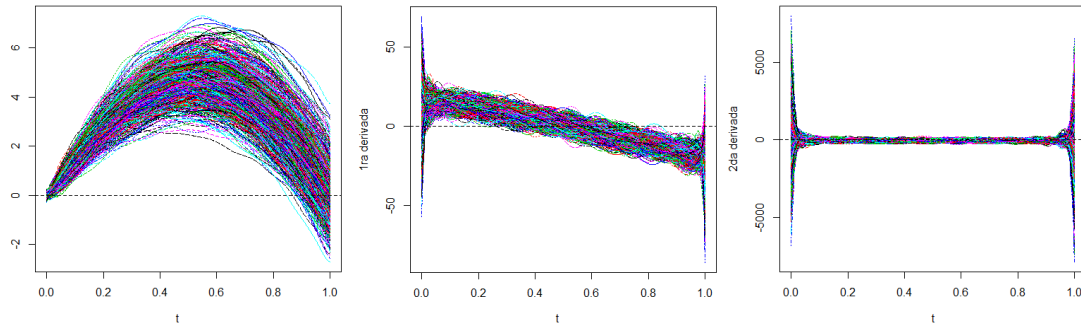


FIGURA 3.2. Funciones estimadas para las observaciones y_i y sus dos primeras derivadas

de control de esta carta, mediante la técnica de bootstrap suavizado. Las funciones se observarán en $n_j = 100$ puntos discretizados equidistantes en el intervalo $I = [0, 1]$. Para

F	T_F^2	T_{MF}^2	fAO	MFHD
5.614608 (0.0476)	14.64009 (0.0953)	14.58031 (0.928)	1.252576 (0.0404)	0.1330322 (0.0065)

TABLA 3.1. Límites de control estimados para las cartas F , T_F^2 , T_{MF}^2 , fAO y MFHD usando bootstrap suavizado, con una tasa de falsas alarmas de 0.005 y su respectiva desviación estándar en paréntesis

el ajuste de todos los perfiles simulados en esta Tesis se realizó suavizamiento spline de orden 6 con 20 bases.

Los límites de control estimados para cada carta se presentan en la Tabla 3.1, los cuales se estimaron para una tasa de falsas alarmas de $\alpha = 0.005$, es decir, que para un conjunto de datos funcionales bajo control, hay un 0.5% de probabilidad que una curva sea identificada como fuera de control dado que no lo está. Para las cartas T_F^2 y T_{MF}^2 se tomaron $K = 4$ componentes principales, teniendo como criterio que al retener entre 4 y 5 componentes, el valor de la diferencia del porcentaje de varianza fue menor a 1%. Como se había mencionado anteriormente, para procesos gaussianos la distribución de la estadística T^2 funcional tiene distribución chi cuadrado con K grados de libertad. Para este caso, $K = 4$ nos da un cuantil $\chi_{4,0.995}^2 = 14.86026$, lo cual representa una ventaja al emplear las cartas T^2 univariada y multivariada, dado que se podrían evaluar las curvas con este límite de control sin emplear bootstrap suavizado u otras técnicas.

Una vez calculados los límites de control de cada carta, se proponen los siguientes esquemas de contaminación en magnitud y forma, empleados en Pan et al. [2019] y Sheu et al. [2013], para evaluar la potencia de las cartas:

- R1. Ruido de magnitud: se aumenta en $\delta_1 \times \|\mu_\chi\|_2$ la media del proceso, para los valores $\delta_1 = 0, 0.02, 0.05, 0.08, 0.11, 0.14, 0.17, 0.2, 0.23, 0.26, 0.29, 0.6, 1, 1.5, 2, 2.5, 3$, como se describe en la ecuación 3.5.

$$y_i^*(t) = \chi(t) + \delta_1 \times \|\mu_\chi\|_2 + \epsilon_i(t), \quad (3.5)$$

donde $\|\mu_\chi\|_2$ es la norma L_2 de la función μ_χ , definida como $\|\mu_\chi\|_2 = \left(\int_0^1 |\mu_\chi(x)|^2 dx \right)^{1/2}$.

- R2. Ruido de forma: se aumenta en $\delta_2\sigma$ la varianza del error ϵ (ecuación 3.6), para los valores $\delta_2 = 0, 1, 2, 3, 4, 5, 10$

$$\epsilon_i \sim N(0, 0.01 + \delta_2\sigma). \quad (3.6)$$

En la gráfica 3.3 a) se puede evidenciar los efectos del ruido R1 clasificados en cuatro niveles de cambios, sobre una muestra de 25 curvas estimadas a partir de las observaciones discretas del proceso. Los cambios son: muy pequeños, cuando el valor $\delta_1 = 0.02, 0.05$; pequeños, cuando el valor $\delta_1 = 0.08, 0.11, 0.14, 0.17, 0.2, 0.23, 0.26, 0.29, 0.6$; moderados, cuando $\delta_1 = 1, 1.5, 2$; grandes, cuando $\delta_1 = 2.5, 3$. Por otro lado, en la gráfica 3.3 b), se puede evidenciar los efectos del ruido R2, los cuales son más difíciles de monitorear dado que, al estimar las curvas fuera de control a partir de observaciones discretas, su comportamiento es muy parecido a los datos bajo control y resultan como motivación a los métodos propuestos en esta Tesis. Para las contaminación R2 se define los siguientes

niveles de cambio: pequeños, cuando el valor $\delta_2 = 0, 1$; moderados, cuando $\delta_2 = 2$; grandes, cuando $\delta_2 = 3, 4$.

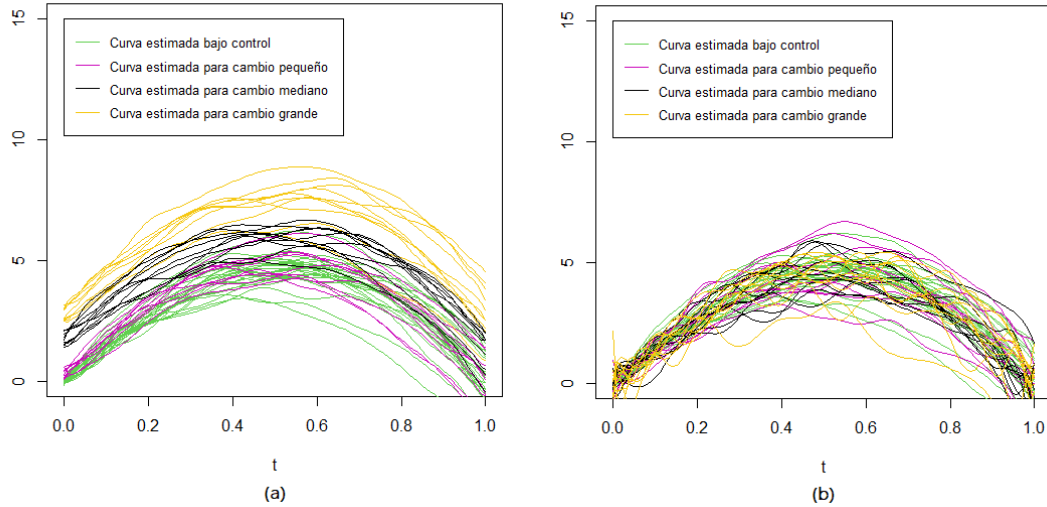


FIGURA 3.3. Muestra de 25 curvas estimadas para las observaciones y_i bajo control y el efecto de la contaminación de magnitud (a) y forma (b). Para las contaminaciones se presentan cambios pequeños, medianos y grandes, de acuerdo a los valores δ_1 y δ_2

Para evaluar la potencia de las cartas se generaron conjuntos de datos totalmente contaminados, bajo los ruidos descritos anteriormente, es decir, desde el tiempo $t = 1$ se empezaron a contaminar los procesos. Cada uno de los vectores de curvas estimados a partir de estos conjuntos, fue comparado con respecto a los vectores de curvas estimados en fase I. Siguiendo la metodología descrita en Sheu et al. [2013], las potencias bajo la contaminación R1 se presentan en la Tabla 3.2.

Como se puede observar en la Tabla 3.2, cuando se presenta una contaminación de magnitud en el modelo, es decir, cuando se afecta la media del proceso estocástico que genera las observaciones, la carta T_F^2 presenta potencias más altas que las demás cuando el cambio es muy pequeño $\delta_1 = 0.02$ o $\delta_1 = 0.05$. Cuando δ_1 toma cambios que no son muy pequeños, las cartas fAO y F son las más potentes, presentando un comportamiento similar con respecto a su potencia. Por ejemplo, para un cambio $\delta_1 = 0.2$, la probabilidad de que la carta fAO detecte un vector de curvas fuera de control, dado que lo está, es 16.28 %.

Hasta este punto se concluye que para un proceso caracterizado por una variable aleatoria funcional, empleando sus primeras dos derivadas con un cambio de magnitud, entre las cartas funcionales multivariadas propuestas, la que resultó más potente en la mayoría de los casos es la carta fAO, presentando desempeños similares a los de la carta F . Se debe tener precaución con el número de componentes principales seleccionadas, ya que estas influyen en los grados de libertad de la distribución de la estadística y una mala elección puede presentar problemas al calcular las falsas alarmas y las potencias en fase I y fase II, respectivamente.

δ_1	Cartas de Control				
	F	T_F^2	T_{MF}^2	fAO	MFHD
0	0.0055 (0.0073)	0.0060 (0.0082)	0.0049 (0.0068)	0.0041 (0.0055)	0.004 (0.0044)
0.02	0.0063 (0.0086)	0.0091 (0.0096)	0.00508 (0.0073)	0.0059 (0.0084)	0.0047 (0.0035)
0.05	0.0105 (0.0102)	0.0112 (0.011)	0.0046 (0.0064)	0.0070 (0.0082)	0.0041 (0.0032)
0.08	0.0182 (0.0126)	0.0156 (0.0127)	0.00458 (0.0065)	0.0148 (0.0118)	0.0043 (0.0039)
0.11	0.0316 (0.0175)	0.0226 (0.0143)	0.005 (0.0072)	0.0337 (0.0202)	0.0049 (0.00311)
0.14	0.0587 (0.0222)	0.0415 (0.0197)	0.0049 (0.007)	0.0559 (0.0225)	0.0042 (0.0046)
0.17	0.0965 (0.0311)	0.0656(0.0265)	0.0043 (0.0068)	0.0994 (0.0313)	0.0051 (0.0047)
0.2	0.1637 (0.0372)	0.0949 (0.03)	0.0044 (0.0065)	0.1628 (0.0366)	0.0047 (0.0043)
0.23	0.2589 (0.0413)	0.1496 (0.0394)	0.0048 (0.0066)	0.2515 (0.042)	0.0047 (0.0041)
0.26	0.3994 (0.0501)	0.2153 (0.0406)	0.0047 (0.0068)	0.3649 (0.0509)	0.0052 (0.00565)
0.29	0.5524 (0.0523)	0.2912 (0.0518)	0.005 (0.006)	0.4878 (0.053)	0.0069 (0.0068)
0.6	1 (0)	0.9826 (0.0142)	0.0048 (0.0064)	0.9989 (0.0037)	0.0073 (0.0069)
1	1 (0)	1 (0)	0.0044 (0.0065)	1 (0)	0.0081 (0.0073)
1.5	1 (0)	1 (0)	0.0046 (0.0064)	1 (0)	0.0081 (0.0075)
2	1 (0)	1 (0)	0.0045 (0.0061)	1 (0)	0.0091 (0.0091)
2.5	1 (0)	1 (0)	0.0044 (0.0065)	1 (0)	0.0094 (0.0091)
3	1 (0)	1 (0)	0.0048 (0.0065)	1 (0)	0.0104 (0.0109)

TABLA 3.2. Potencia promedio estimada de las cartas F , T_F^2 , T_{MF}^2 , fAO y MFHD con su respectiva desviación estándar en paréntesis, cuando el proceso es contaminado con el tipo de ruido R1

δ_2	F	T_F^2	T_{MF}^2	fAO	MFHD
0	0.0058 (0.0076)	0.0068 (0.0083)	0.00504 (0.0068)	0.0048 (0.007)	0.0044 (0.00492)
1	0.0061 (0.0076)	0.0086 (0.0103)	0.6019 (0.0497)	0.3474 (0.0449)	0.0049 (0.0047)
2	0.0076 (0.0081)	0.0128 (0.0093)	0.7835 (0.0404)	0.7781 (0.0427)	0.0041 (0.0051)
3	0.0094 (0.0098)	0.0165 (0.0118)	0.8631 (0.0332)	0.9307 (0.0251)	0.0042 (0.0049)
4	0.0128 (0.0105)	0.0207 (0.0134)	0.9061 (0.029)	0.9732 (0.0159)	0.004 (0.0051)

TABLA 3.3. Potencia promedio estimada de las cartas F , T_F^2 , T_{MF}^2 , fAO y MFHD con su respectiva desviación estándar en paréntesis, cuando el proceso es contaminado con el tipo de ruido R2

En la gráfica 3.4 se puede observar una corrida de las cartas propuestas. Se presentan las cartas de control en fase II, bajo la contaminación de forma descrita en R2 con $\delta_2 = 0.2$, para 25 curvas nuevas, contaminadas desde el tiempo $t = 1$, es decir, desde la primera corrida. Se observa que, para esta corrida en específico, las cartas más potentes fueron la la fAO y la T_{MF}^2 , dado que detectaron el primer cambio en el proceso en la primera corrida (fAO) y en la cuarta (T_{MF}^2), respectivamente.

Las potencias para bajo ruido R2 se presentan en la Tabla 3.3. En cuanto al cambio de forma, se puede observar que para $\delta_2 = 1$ la carta T_{MF}^2 fue más potente que las demás, es decir, identifica mejor los cambios de forma pequeños, comparada con las demás cartas. Sin embargo, para $\delta_2 = 2$ la carta T_{MF}^2 y fAO tuvieron potencias similares y, para valores de $\delta_2 > 3$, se observa que la carta fAO es la más potente de las evaluadas, indicando que en la medida que una contaminación de forma al proceso vaya aumentando, la carta fAO aumenta su potencia y es la más indicada para detectar estos cambios.

Las cartas univariadas no presentaron potencias altas, evidenciando que los cambios de forma se identificaron mejor empleando las derivadas del proceso, es decir, empleando

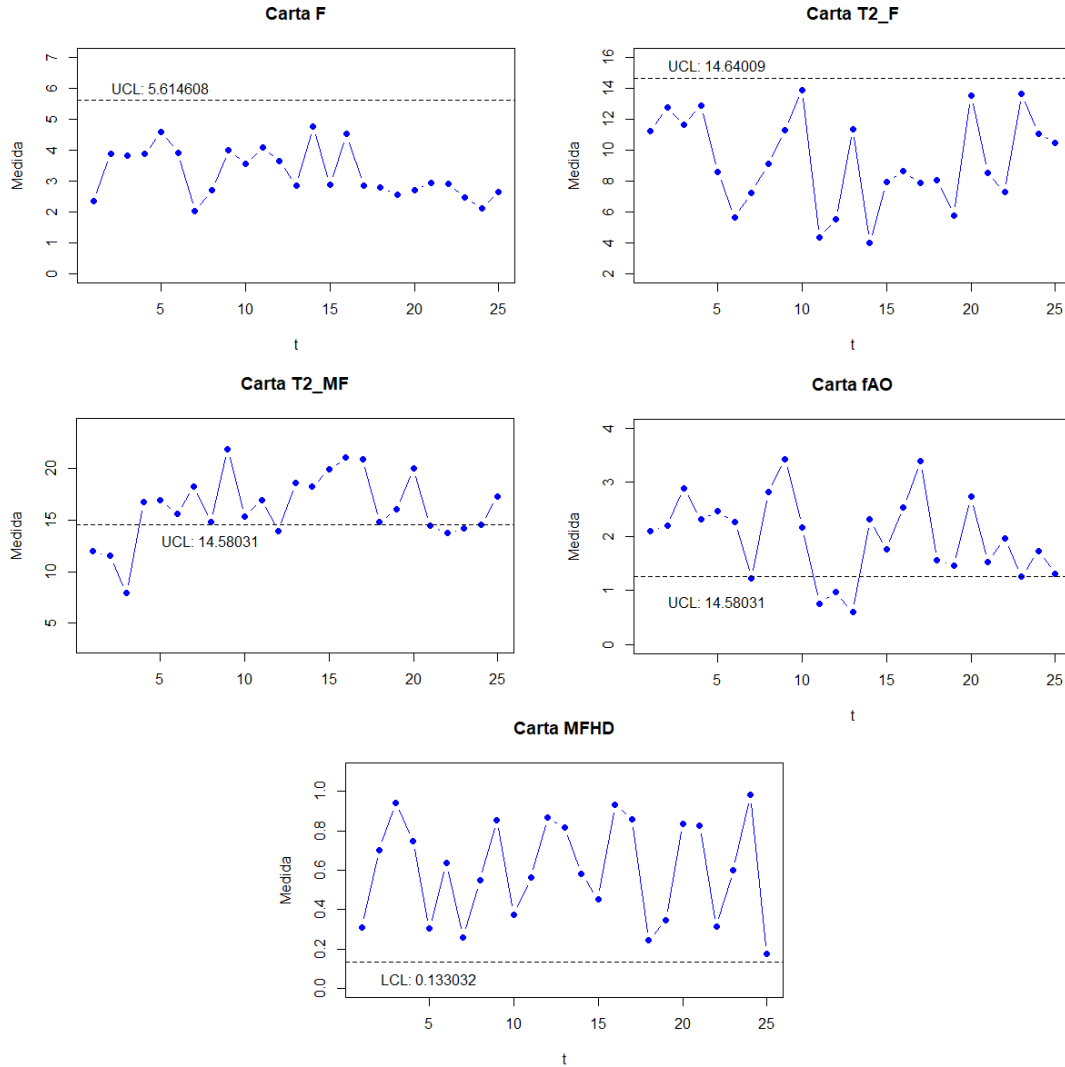


FIGURA 3.4. Corrida de una carta de control para las propuestas F , T_F^2 , T_{MF}^2 , fAO y MFHD en fase II para 25 observaciones nuevas, bajo la contaminación R2, con $\delta_2 = 0.2$

más información. Otro hecho que cabe resaltar es que las cartas multivariadas empezaron a identificar cambios muy pequeños rápidamente.

En conclusión, al emplear las derivadas para incluir mas información a un proceso funcional univariado, se puede identificar observaciones fuera de control, de forma más potente, si los cambios son de forma. A pesar de que los cambios de magnitud (Tabla 3.2) se identifican rápidamente sin emplear las derivadas, se observa que emplear la carta fAO (que emplea la atipicidad ajustada funcional) también presenta potencias altas (en un par de casos mayor) que la carta F.

En cuanto a tiempos de ejecución, los cálculos fueron realizados en el software R, y en la tabla 3.4 se reporta el tiempo que se demora el procesamiento de cada una de las

Carta	Tiempo de cómputo
F	0.0009968281 secs
T_F^2	1.11241 secs
fAO	25.55533 secs
MFHD	23.60429 secs
T_{MF}^2	2.211041 secs

TABLA 3.4. Tiempos de ejecución de las estadísticas asociadas a las cartas F y T_F^2 para una curva estimada y T_{MF}^2 , fAO y MFHD para un vector de 3 curvas, donde cada una de ellas se encuentra observada en 100 puntos discretos del tiempo

estadísticas propuestas en esta Tesis, compiladas en un procesador Intel Core i7 de séptima generación, 7600U 2.8Ghz, doble núcleo. La estadística F fue calculada para una curva con respecto a la curva media y la desviación estándar histórica estimada, mientras que la estadística T_F^2 fue calculada para una curva con respecto al conjunto histórico de 1000 curvas. Las estadísticas de las cartas multivariadas fAO, MFHD y T_{MF}^2 fueron calculadas todas con respecto al conjunto histórico de 1000 vectores de curvas.

3.3. Comparación cartas multivariadas

En esta sección se presenta la simulación de un dato funcional multivariado empleando la carta de control propuesta en Pan et al. [2019], la cual es una carta multivariada no paramétrica EWMA de rango espacial basada en un modelo de regresión de vectores de soporte (SVM), para comparar las cartas propuestas en esta Tesis.

El objetivo será estimar en fase I, donde todos los datos funcionales multivariados estén bajo control, los parámetros requeridos (entre ellos los límites de control) y luego, en una segunda fase, evaluar el rendimiento para evidenciar qué carta presenta mejor comportamiento. El dato funcional multivariado se genera siguiendo el procedimiento planteado por Pan et al. [2019], descrito en las siguientes ecuaciones, en el cual se observan muestras aleatorias de tamaño n_j para un perfil no lineal multivariado recolectadas en 162 puntos discretos de tiempo $j = 1, 2, \dots$ con $i = 1, 2, \dots, n_j$.

$$\begin{aligned}
y_{ij1} &= \theta_{11} * (1 - \theta_{12} * e^{-0.036*x_{ij}}) + \frac{258.556 - \theta_{11}}{1 + e^{0.023*(x_{ij}-313.350)}} + \epsilon_{ij1}, \\
y_{ij2} &= 257.930 * (1 - 0.052 * e^{-0.048*x_{ij}}) + \frac{260.361 - 257.930}{1 + e^{0.021*(x_{ij}-294.988)}} + \epsilon_{ij2}, \\
y_{ij3} &= \theta_{31} * (1 - 0.060 * e^{-0.046*x_{ij}}) + \frac{(261.192 - \theta_{31})}{1 + e^{0.022*(x_{ij}-296.595)}} + \epsilon_{ij3}, \\
y_{ij4} &= 258.932 * (1 - 0.054 * e^{-0.050*x_{ij}}) + \frac{(261.697 - 258.932)}{1 + e^{0.019*(x_{ij}-280.978)}} + \epsilon_{ij4},
\end{aligned} \tag{3.7}$$

donde $\theta_{11} = 256.748$, $\theta_{12} = 0.058$, $\theta_{13} = 259.074$. El vector de errores $\epsilon = (\epsilon_{ij1}, \epsilon_{ij2}, \epsilon_{ij3}, \epsilon_{ij4})$ es un vector aleatorio, con distribuciones de probabilidad multivariadas

centradas en el vector $\mathbf{0} = (0, 0, 0, 0)$ y matrix de varianzas y covarianzas Σ definida como:

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 & \rho\sigma_1\sigma_3 & \rho\sigma_1\sigma_4 \\ \rho\sigma_2\sigma_1 & \sigma_2^2 & \rho\sigma_2\sigma_3 & \rho\sigma_2\sigma_4 \\ \rho\sigma_3\sigma_1 & \rho\sigma_3\sigma_2 & \sigma_3^2 & \rho\sigma_3\sigma_4 \\ \rho\sigma_4\sigma_1 & \rho\sigma_4\sigma_2 & \rho\sigma_4\sigma_3 & \sigma_4^2 \end{bmatrix}, \quad (3.8)$$

la cual define la estructura de covarianza *entre* los perfiles.

Para las simulaciones de los perfiles no lineales multivariados $\{(x_{ij}, y_{ijk}); i = 1, 2, \dots, n_j, j = 1, 2, \dots, k = 1, 2, 3, 4\}$ se tomará $x_{.j} = (x_{1j}, x_{2j}, \dots, x_{n_jj})$ como el vector $(0, 3, 6, \dots, 483)$, de acuerdo a los valores tomados en Pan et al. [2019]. Se tomará $n_j = 162$, ϵ bajo distribución normal multivariada $NM_4(\mathbf{0}, \Sigma)$ y distribución exponencial multivariada $MVE_4(\mathbf{0}, \Sigma)$, $\rho \in \{0.1, 0.5, 0.9\}$, indicando correlaciones baja, media y alta entre perfiles, $\sigma_1 = 1$, $\sigma_2 = 1$, $\sigma_3 = 1$, $\sigma_4 = 1$. En fase I se realizarán simulaciones de 500 perfiles no lineales multivariados para la estimación de la carta, con el objetivo de estimar los límites de control, bajo una tasa de falsas alarmas $\alpha = 0.005$, que permitan una correcta identificación en fase II de los procesos fuera de control. En el artículo ellos emplean muestras históricas pequeñas con 20 y 40 curvas, sin embargo las cartas propuestas en esta Tesis requieren tamaños de muestra grandes en fase I para la correcta identificación en fase II de los procesos fuera de control.

Un ejemplo de una observación del perfil no lineal multivariado generado por este modelo se puede visualizar en la gráfica 3.5. Se generarán 500 conjuntos de $m = 500$ perfil no lineales multivariados en fase I, para determinar los límites de control en cada carta mediante la técnica de bootstrap suavizado y las funciones se observarán en 162 puntos discretizados equidistantes $x_{.j} = (0, 3, 6, \dots, 483)$. Para el ajuste de todos los perfiles simulados en esta Tesis se realizó suavizamiento spline de orden 6 con 20 bases, con el objetivo de que las curvas estimadas presenten un buen ajuste sin tener problemas de sobreajuste, evaluado la calidad de la estimación mediante el error cuadrado medio [Ramsay and Silverman, 2005]. En la gráfica 3.6 se puede observar el conjunto de datos suavizado. Los límites de control para cada carta se estimaran para una tasa de falsas alarmas de $\alpha = 0.005$, es decir, que para un conjunto de datos funcionales multivariados bajo control, hay un 0.5% de probabilidad que una curva sea identificada como fuera de control dado que no lo está. Para la cartas T^2 funcional multivariada se tomó $K = 5$ componentes principales después de evaluar criterios de diferencia de porcentaje de varianza retenido, tomando valores cercanos al 1%, de forma análoga a la realizada con un proceso y sus dos primeras derivadas. Los límites de control estimados para cada una de las cartas propuestas y sus desviaciones estándar se presentan en la Tabla 3.5.

Se presentan dos metodologías para evaluar el rendimiento de la carta. En la primera, se empleará la metodología implementada por Sheu et al. [2013], en la cual se calculan las potencias de las cartas, simulando conjuntos de datos totalmente contaminados con un número finito de observaciones, calculando la proporción de curvas que se identifican como fuera de control cuando efectivamente lo están. La otra metodología será calculando el ARL de las curvas, donde se contaminaran a partir del primer tiempo, $t = 1$, e identificando cuantos vectores de curvas necesita cada una de las cartas propuestas para

Carta	ρ	Distribución Σ	Límite de control	Desviación estándar
T_{MF}^2	0.1	Normal multivariada	15.65398	1.165398
T_{MF}^2	0.5	Normal multivariada	15.97254	0.9967762
T_{MF}^2	0.9	Normal multivariada	16.4292	1.7088
T_{MF}^2	0.9	Exponencial multivariada	15.45104	1.0285106
fAO	0.1	Normal multivariada	1.001349	0.01081425
fAO	0.5	Normal multivariada	0.9975891	0.01061709
fAO	0.9	Normal multivariada	1.001688	0.009988883
fAO	0.9	Exponencial multivariada	0.9998621	0.01029016
MFHD	0.1	Normal multivariada	0.5139735	0.002521893
MFHD	0.5	Normal multivariada	0.5139849	0.002524632
MFHD	0.9	Normal multivariada	0.5144685	0.00265884
MFHD	0.9	Exponencial multivariada	0.5135105	0.002345598

TABLA 3.5. Límites de control estimados para las cartas T_{MF}^2 , fAO y MFHD usando bootstrap suavizado con una tasa de falsas alarmas de 0.005 con su respectiva desviación estándar

empezar a detectar fallas en el proceso.

Se proponen los siguientes esquemas de contaminación en magnitud y forma empleados en Pan et al. [2019] para evaluar la potencia de las cartas:

- C1. θ_{11} cambiará por $\delta_{11}\sigma_1 + \theta_{11}$ y se evaluarán los valores $\rho = \{0.1, 0.5, 0.9\}$ bajo la distribución normal multivariada para ϵ . δ_{11} tomará los valores $\{0, 0.4, 0.6, 0.8, 1\}$.
- C2. θ_{12} cambiará por $\delta_{12}\sigma_1 + \theta_{12}$ y se evaluarán el valor $\rho = 0.9$ bajo las distribuciones normal multivariada y exponencial multivariada para ϵ . δ_{12} tomará los valores $\{0, 0.004, 0.006, 0.008, 0.015\}$.
- C3. σ_1 cambiará por $\gamma_1\sigma_1$ y σ_3 cambiará por $\gamma_3\sigma_3$, y se evaluará el valor $\rho = 0.9$ bajo la distribución normal multivariada para ϵ . γ_1 tomará valores $\{0.96, 0.94, 0.92, 0.9\}$ y γ_3 tomará valores $\{1.04, 1.06, 1.08, 1.1\}$.

En la gráfica 3.7 se puede evidenciar el efecto del ruido C1 en el proceso y_{ij1} para $\delta_1 = 0.6$ y el efecto del ruido C2 en el proceso y_{ij2} para $\delta_2 = 0.01$.

En la gráfica 3.8 se puede evidenciar el efecto del ruido C3 en los procesos y_{ij3} y y_{ij4} para $\gamma_1 = 0.94$ y $\gamma_3 = 1.06$.

Para evaluar la potencia de las cartas se emplearon $m_3 = 162$ observaciones del dato funcional en fase II, las cuales se monitorean individualmente, bajo las contaminaciones C1, C2 y C3 las cuales se presentan en las Tablas 3.6, 3.7 y 3.8. Las cartas a evaluar son la T_{MF}^2 , fAO y MFHD, comparadas con la carta SVM propuesta en [Pan et al., 2019], empleando *vectores de soporte (SVM)*.

En la Tabla 3.6 se puede observar que la carta fAO presenta las potencias más altas cuando se varía δ_{11} de acuerdo a la contaminación C1. Este comportamiento se presenta

δ_{11}	ρ	SVM	T_{MF}^2	fAO	MFHD
0	0.1	0.005	0.020 (0.025)	0.007 (0.007)	0.006 (0.005)
0	0.5	0.005	0.009 (0.016)	0.005 (0.005)	0.005 (0.006)
0	0.9	0.005	0.006 (0.007)	0.006 (0.007)	0.006 (0.007)
0.4	0.1	0.008	0.007 (0.021)	0.028 (0.013)	0.006 (0.006)
0.4	0.5	0.008	0.008 (0.020)	0.042 (0.016)	0.006 (0.005)
0.4	0.9	0.008	0.008 (0.006)	0.23 (0.04)	0.007 (0.005)
0.6	0.1	0.024	0.009 (0.030)	0.09 (0.022)	0.006 (0.005)
0.6	0.5	0.026	0.008 (0.019)	0.189 (0.03)	0.006 (0.004)
0.6	0.9	0.046	0.009 (0.004)	0.299 (0.03)	0.007 (0.005)
0.8	0.1	0.073	0.01 (0.012)	0.26 (0.035)	0.006 (0.005)
0.8	0.5	0.076	0.015 (0.010)	0.29 (0.04)	0.007 (0.006)
0.8	0.9	0.095	0.02 (0.005)	0.42 (0.01)	0.0075 (0.004)
1	0.1	0.103	0.015 (0.011)	0.549 (0.039)	0.007 (0.005)
1	0.5	0.105	0.02 (0.01)	0.869 (0.024)	0.007 (0.005)
1	0.9	0.120	0.1 (0.005)	0.989 (0.009)	0.009 (0.005)

TABLA 3.6. Potencia de las cartas de control *SVM* (sin desviación estándar), T_{MF}^2 , fAO y MFHD, para los perfiles generados por las ecuaciones (3.7), con error que se distribuye normal $N_4(\mathbf{0}, \Sigma)$, Σ generada por la ecuación (3.8), con valores de correlación bajo ($\rho = 0.1$), medio ($\rho = 0.5$) y alto ($\rho = 0.9$), bajo el esquema de contaminación C1 con valores δ_{11} entre 0 y 1.5 y sus respectivas desviaciones estándar en paréntesis

para todos los valores tanto de δ_{11} como de ρ , es decir, es la carta más potente entre las desarrolladas para correlación baja, media y alta entre los procesos del dato funcional multivariado. Por ejemplo, para una contaminación $\delta_{11} = 0.4$, se puede observar en la Tabla 3.6 que la probabilidad de marcar un vector de curvas fuera de control, dado que efectivamente lo está, es del 23% cuando la correlación es $\rho = 0.9$ (alta). Para el caso de correlación moderada, $\rho = 0.5$, la probabilidad de detectar un vector de curvas fuera de control, dado que lo está es de 4.2%, y cuando la correlación es baja, es decir $\rho = 0.1$ esta probabilidad fue de 2.8%. En general, para todas las cartas, a medida que disminuye la correlación entre los perfiles también disminuye la potencia. Esto puede estar relacionado con que la mayoría de cambios se realizaron sobre una sola curva del vector de curvas y, al presentar altas correlaciones entre observaciones (bajo la estructura de la matriz Σ 3.8), esto facilita el monitoreo de los perfiles.

La carta *SVM* fue la segunda, después de la carta fAO, en presentar mejores resultados para identificar procesos con este cambio descrito en C1. Después de estas cartas, tenemos la carta T_{MF}^2 , que presentó potencias más pequeñas que las otras propuestas en esta Tesis, pero aun así presentó mejores valores que la carta MFHD, la cual no identificó adecuadamente los perfiles multivariados fuera de control. Respecto al cambio en las correlaciones entre procesos, se puede observar que para todas las cartas la potencia aumentó cuando los procesos estaban **altamente** correlacionados, indicando que las cartas empleadas para procesos que se monitoreen empleando perfiles multivariados mediante un enfoque de datos funcionales, presentarán potencias más altas cuando los perfiles tengan correlaciones altas entre ellos. Los procesos con correlaciones bajas se podrán monitorear bajo los esquemas propuestos, sin embargo la identificación de las observaciones fuera de control tomará más tiempo para ser detectadas.

δ_{12}	Σ	SVM	T_{MF}^2	fAO	MFHD
0.000	N_4	0.005	0.006 (0.007)	0.006 (0.007)	0.006 (0.007)
0.000	MVE	0.005	0.013 (0.013)	0.006 (0.005)	0.004 (0.005)
0.004	N_4	0.005	0.008 (0.007)	0.011 (0.004)	0.006 (0.008)
0.004	MVE	0.007	0.009 (0.013)	0.012 (0.036)	0.005 (0.006)
0.006	N_4	0.008	0.008 (0.007)	0.023 (0.005)	0.008 (0.007)
0.006	MVE	0.016	0.01 (0.012)	0.017 (0.008)	0.007 (0.006)
0.008	N_4	0.027	0.008 (0.008)	0.035 (0.009)	0.009 (0.006)
0.008	MVE	0.047	0.011 (0.014)	0.028 (0.012)	0.007 (0.006)
0.015	N_4	0.117	0.009 (0.007)	1 (0)	0.009 (0.004)
0.015	MVE	0.117	0.010 (0.013)	1 (0)	0.007 (0.007)

TABLA 3.7. Potencia de las cartas de control SVM (sin desviación estándar), T_{MF}^2 , fAO y MFHD, para los perfiles generados por las ecuaciones (3.7), con error que se distribuye normal $N_4(\mathbf{0}, \Sigma)$ y exponencial multivariada $MVE(\Sigma)$, Σ generada por la ecuación (3.8), bajo el esquema de contaminación C2 con valores δ_{12} entre 0 y 0.02 y sus respectivas desviaciones estándar en paréntesis

En la Tabla 3.7 se puede observar que la carta fAO es la más potente respecto a las demás propuestas, cuando se emplea la contaminación C2. En este caso, se compararon las cartas cambiando la estructura de covarianza entre los procesos, todos con una correlación alta, $\rho = 0.9$. Cuando $\delta_{12} = 0.004$ la carta fAO identifica correctamente un 1.1% de las curvas fuera de control, evidenciando mayor efectividad con respecto a las demás cartas. Sin embargo, como tendencia general, las cartas fueron ligeramente más potentes bajo la distribución normal multivariada que respecto a la distribución exponencial multivariada (MVE), concluyendo que cuando los datos funcionales multivariados provengan de procesos gaussianos, se presentarán potencias más altas en las cartas propuestas. Se puede observar que, independientemente a la distribución de los errores, se pueden emplear las cartas propuestas para monitoreo de perfiles no lineales multivariados, así como también se observa que a medida que el valor δ_{12} aumenta, también aumentan las potencias de las cartas.

En la Tabla 3.8 se puede evidenciar que la carta SVM presentó potencias más altas, en general, que las cartas propuestas para la contaminación C3. Para este caso se emplearon valores altos de correlación ($\rho = 0.9$) y con error que se distribuye normal $N_4(\mathbf{0}, \Sigma)$, Σ generada por la ecuación (3.8). En general, la carta fAO en este caso tuvo potencias muy bajas mientras que la carta T^2 funcional multivariada fue la que presentó mejores resultados entre las propuestas en la Tesis. En este tipo de contaminación se puede observar que, en general, los cambios en γ_1 se identificaron más rápidamente que los cambios en γ_3 , sin embargo, los cambios de este último son mucho más pequeños que γ_1 . En conclusión, para las cartas propuestas será muy difícil identificar cambios muy pequeños que se presenten en la estructura de varianza y covarianza entre los datos funcionales univariados marginales que componen el dato funcional multivariado.

Para la segunda metodología se proponen los siguientes esquemas de contaminación en magnitud y forma, empleados en Pan et al. [2019] para calcular el ARL de las cartas:

γ_3	γ_1	SVM	T_{MF}^2	fAO	MFHD
1.04	0.96	0.052	0.022 (0.012)	0.007 (0.007)	0.008 (0.007)
1.04	0.94	0.079	0.024 (0.012)	0.008 (0.007)	0.006 (0.006)
1.04	0.92	0.097	0.022 (0.011)	0.006 (0.006)	0.006 (0.007)
1.04	0.9	0.11	0.021 (0.012)	0.006 (0.005)	0.004 (0.005)
1.06	0.96	0.076	0.022 (0.013)	0.010 (0.007)	0.008 (0.008)
1.06	0.94	0.088	0.024 (0.013)	0.010 (0.007)	0.008 (0.007)
1.06	0.92	0.101	0.024 (0.013)	0.009 (0.008)	0.007 (0.007)
1.06	0.9	0.11	0.022 (0.011)	0.009 (0.007)	0.007 (0.006)
1.08	0.96	0.093	0.024 (0.011)	0.013 (0.01)	0.011 (0.008)
1.08	0.94	0.099	0.026 (0.011)	0.012 (0.009)	0.010 (0.007)
1.08	0.92	0.106	0.024 (0.012)	0.011 (0.008)	0.008 (0.008)
1.08	0.9	0.113	0.022 (0.011)	0.012 (0.009)	0.009 (0.008)
1.1	0.96	0.106	0.027 (0.014)	0.017 (0.01)	0.012 (0.009)
1.1	0.94	0.108	0.026 (0.012)	0.016 (0.01)	0.012 (0.008)
1.1	0.92	0.112	0.024 (0.013)	0.015 (0.01)	0.011 (0.007)
1.1	0.9	0.116	0.024 (0.013)	0.017 (0.011)	0.012 (0.008)

TABLA 3.8. Potencia de las cartas de control *SVM* (sin desviación estándar), T_{MF}^2 , fAO y MFHD, para los perfiles generados por las ecuaciones (3.7), con error que se distribuye normal $N_4(\mathbf{0}, \mathbf{\Sigma})$, $\mathbf{\Sigma}$ generada por la ecuación (3.8) y correlación alta ($\rho = 0.9$) entre los procesos, bajo el esquema de contaminación C3 y sus respectivas desviaciones estándar en paréntesis

F1. θ_{11} cambiará por $\delta_{11}\sigma_1 + \theta_{11}$, donde δ_{11} tomará los valores $\{0, 0.4, 0.8\}$.

F2. θ_{12} cambiará por $\delta_{12}\sigma_1 + \theta_{12}$, donde δ_{12} tomará los valores $\{0.004, 0.008, 0.015\}$,

para los perfiles generados por las ecuaciones (3.7), con error que se distribuye normal $N_4(\mathbf{0}, \mathbf{\Sigma})$, $\mathbf{\Sigma}$ generada por la ecuación (3.8) y correlación alta ($\rho = 0.9$).

En la gráfica 3.7 se puede evidenciar el efecto de los ruidos F1 y F2 (dado que son los mismos que C1 y C2) en los procesos y_{ij1} y y_{ij2} , respectivamente. Estos ruidos cambian en forma las observaciones del proceso, dado que la media del proceso se mantiene pero se genera una mayor variabilidad entre las observaciones, a medida que los valores δ_1 y δ_2 aumentan.

Para esta simulación se tendrá en cuenta la metodología de Pan et al. [2019], en la cual se calcula el ARL de las cartas propuestas. El valor de los ARL y sus respectivas desviaciones estándar, bajo las contaminaciones F1 y F2, se presentan en las Tablas 3.6 y 3.7. Las cartas a evaluar son la T_{MF}^2 , fAO, MFHD y se comparan con la carta *SVM* que emplea vectores de soporte.

En la Tabla 3.9 se puede observar que la carta fAO presenta la longitud promedio de corrida más alta cuando se varía δ_{11} de acuerdo a la contaminación F1. Este comportamiento se presenta para todos los valores tanto de δ_{11} , es decir, es la carta más potente

δ_{11}	SVM	fAO	MFHD	T_{MF}^2
0	201.092 (188.381)	196.093 (197.165)	196.677 (191.4338)	194.1382 (201.31677)
0.4	71.577 (104.542)	37.088 (38.62)	159.479 (144.8004)	148.3669 (147.9090)
0.8	8.957 (1.729)	4.04 (3.25)	146.2733 (141.3619)	49.7545 (48.2037)

TABLA 3.9. ARL promedio estimado de las cartas *SVM*, fAO, MFHD y T_{MF}^2 con su respectiva desviación estándar en paréntesis, cuando el proceso es contaminado con el tipo de ruido F1

δ_{12}	SVM	fAO	MFHD	T_{MF}^2
0.004	198.867 (193.226)	90.276 (83.192)	152.7722 (145.0024)	190.4549 (188.4541)
0.008	36.798 (67.527)	28.4156 (27.9597)	146.5702 (138.7390)	149.82 (148.7755)
0.015	8.513 (1.331)	4.562 (3.9620)	144.4833 (131.5831)	96.9915 (89.9351)

TABLA 3.10. ARL promedio estimado de las cartas *SVM*, fAO, MFHD y T_{MF}^2 con su respectiva desviación estándar en paréntesis, cuando el proceso es contaminado con el tipo de ruido F2

entre las desarrolladas y la propuesta por Pan et al. [2019]. También se puede observar que la carta *SVM* fue, después de la fAO, la que presentó mejores longitudes promedio de corrida, aunque para cambios pequeños la diferencia fue muy alta con la propuesta en esta Tesis. Por otro lado, las cartas MFHD y T_{MF}^2 fueron las que presentaron menores longitudes promedio de corrida.

En la Tabla 3.10 se puede observar que, nuevamente, la carta fAO tiene mejor rendimiento (medido con la longitud promedio de corrida), respecto a las demás propuestas. En este caso se puede observar que las cartas presentan rendimientos altos para valores pequeños de δ_{12} , ya que en el primer cambio tomado de 0 a 0.004 la carta fAO necesita 90 vectores de curvas aproximadamente para identificar un cambio mientras, por ejemplo, la carta *SVM* necesita casi 200, la cual es cercana a su tasa de falsas alarmas. Vale aclarar que en la Tabla 3.10 no se simuló $\delta_{12} = 0$ dado que esto lleva al mismo modelo en el que $\delta_{11} = 0$, cuyo resultado está en la tabla 3.9 .

En conclusión, para una correlación alta ($\rho = 0.9$) y bajo una distribución gaussiana de los errores asociados a los procesos que general los vectores de curvas, una de las mejores cartas resulta empleando la atipicidad ajustada funcional (fAO) como estadística para monitoreo de perfiles no lineales multivariados.

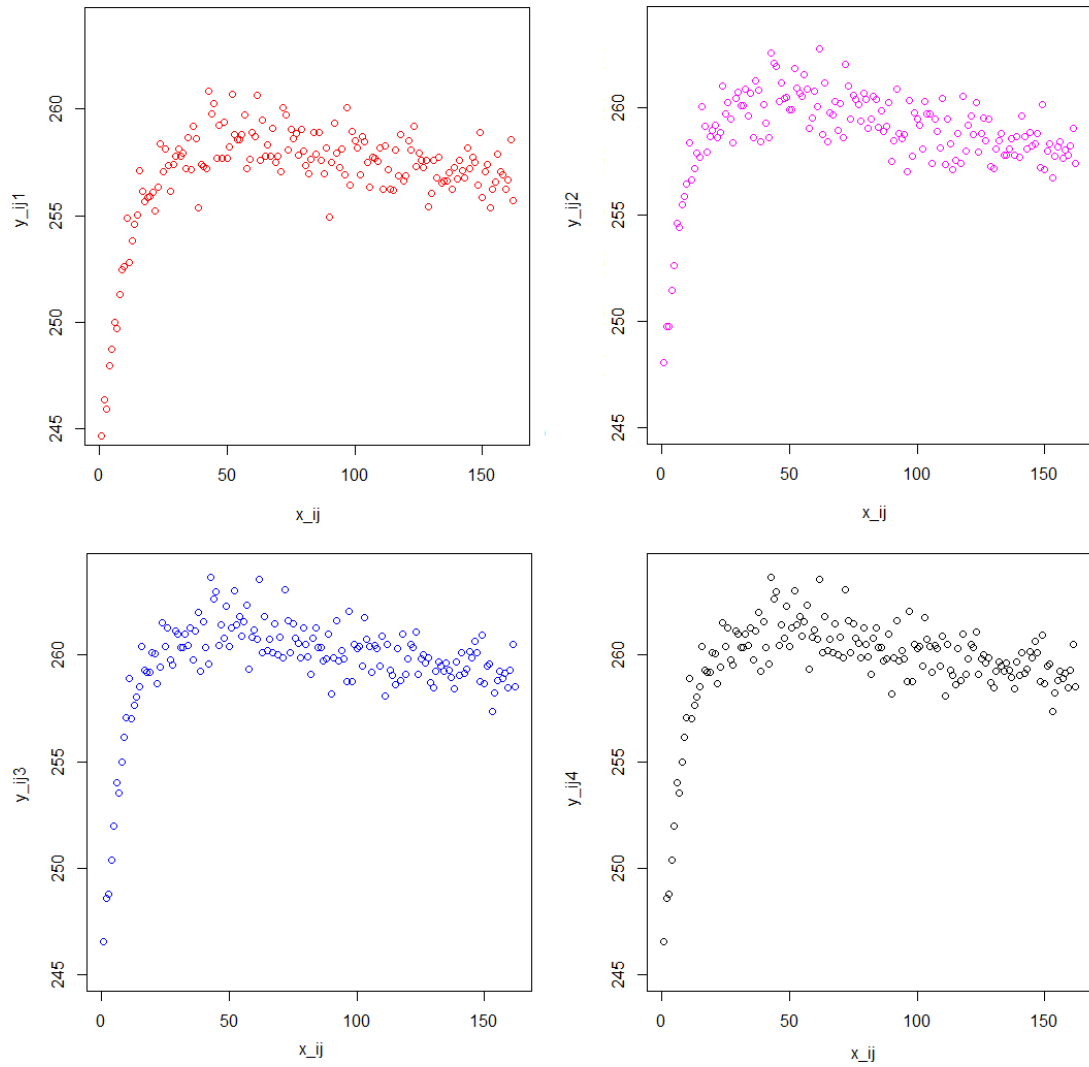


FIGURA 3.5. Realización del perfil no lineal multivariado especificado por las ecuaciones (3.7). La fila superior corresponde a los dos primeros procesos del perfil, y la fila inferior corresponde a los dos últimos

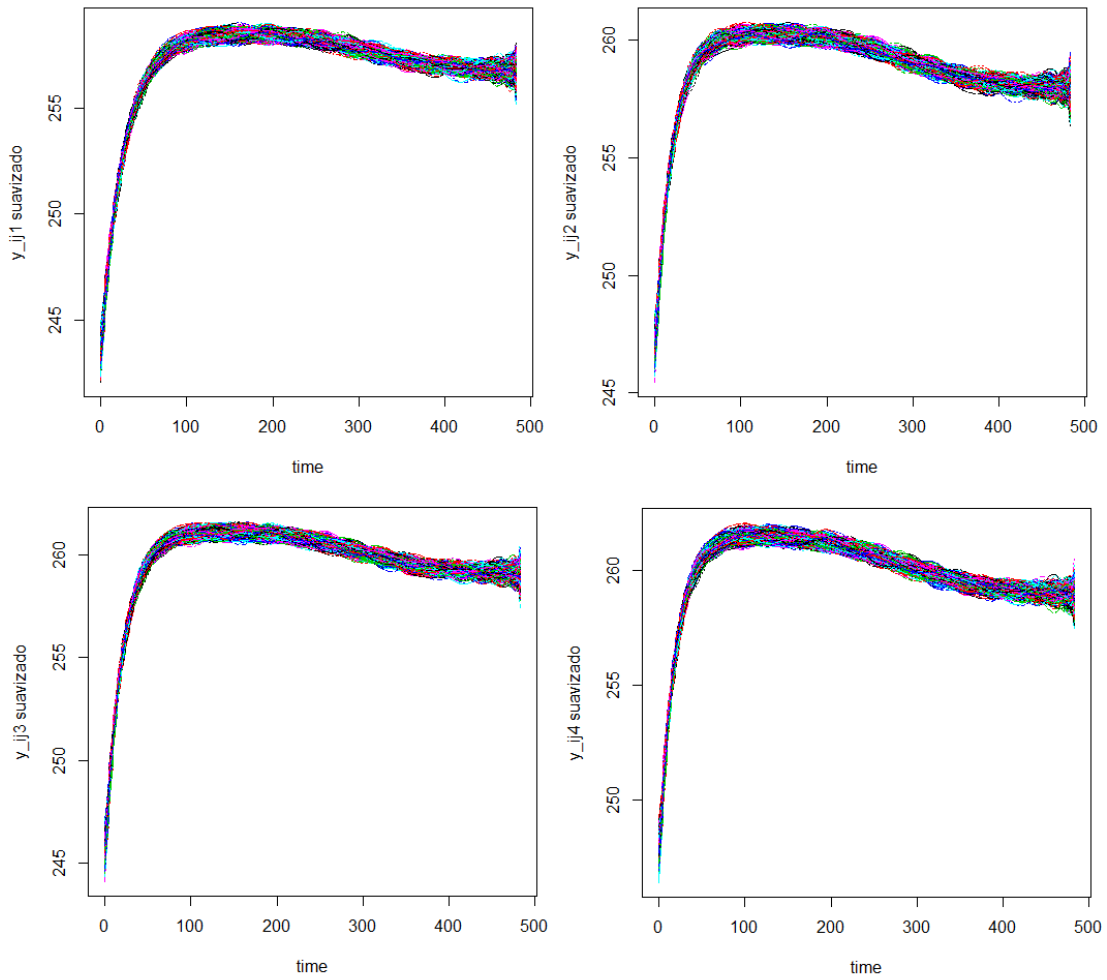


FIGURA 3.6. Suavizado del perfil no lineal multivariado especificado por las ecuaciones (3.7). La fila superior corresponde a los dos primeros procesos del perfil, y la fila inferior corresponde a los dos últimos

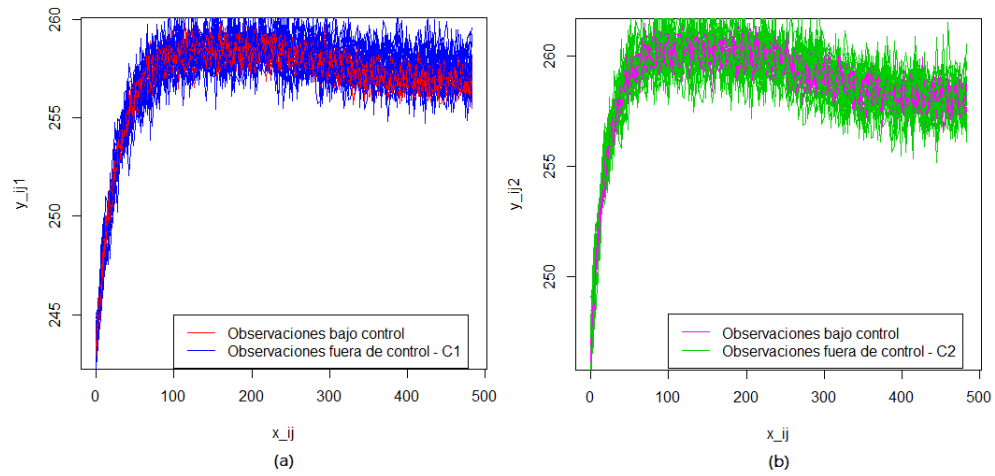


FIGURA 3.7. Contaminación para los dos primeros procesos del perfil no lineal multivariado, descritos en la ecuación 3.8, bajo los esquemas C1 con $\delta_1 = 0.6$ (a) y C2 con $\delta_2 = 0.01$ (b), respectivamente

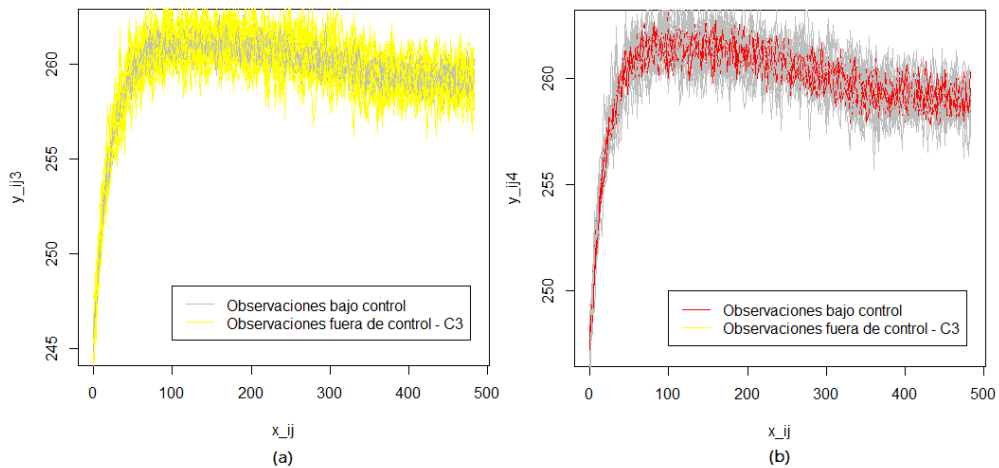


FIGURA 3.8. Contaminación para los procesos los dos últimos procesos del perfil no lineal multivariado, descritos en la ecuación 3.8, bajo el esquema C3 con $\gamma_1 = 0.94$ (a) y $\gamma_3 = 1.06$ (b)

Aplicación a datos reales

En esta sección se realizará una aplicación de las cartas propuestas, sobre un conjunto de datos relativos a la fluorescencia en la producción de azúcar [Munck et al., 1998], por medio de pruebas de espectroscopía. La fluorescencia es la propiedad de algunos átomos y moléculas de absorber luz en una longitud de onda en particular (excitación) y después emitir luces de mayor longitud de onda (emisión) después de un breve intervalo.

Esta técnica se presenta generalmente como espectros de emisión, que se visualizan como un gráfico de la intensidad de fluorescencia frente a la longitud de onda (en nanómetros) para una longitud de onda de excitación. Los espectros son series de picos, o líneas, superpuestas al ruido, donde cada pico surge de una característica de absorción o un compuesto característico [Guevara and Vargas, 2016]. En la investigación realizada por Munck et al. [1998], el azúcar se muestreó continuamente durante 8 horas para hacer una media representante de la muestra para un turno (periodo de 8 horas). Se tomaron muestras durante los tres meses de operación a finales de otoño de una planta de azúcar en Escandinavia, dando un total de 268 muestras. El azúcar se muestreó directamente de la operación unitaria final (centrifugadora) del proceso. El azúcar se disolvió en agua no tamponada (2,25 g / 15 ml) y la solución se midió espectrofluorométricamente en una cubeta de 10 mm por 10 mm en un espectrofluorómetro PE LS50B. Los datos sin suavizar se obtuvieron del fluorómetro. Para cada muestra, los espectros de emisión de 275 a 560 nm se midieron en intervalos de 0,5 nm (571 longitudes de onda) a siete longitudes de onda de excitación: 230, 240, 255, 290, 305, 325, 340 nm [Guevara and Vargas, 2016].

Para la aplicación de las cartas propuestas en la Tesis, se evalúan solamente las longitudes correspondientes a 230, 240, 255 y 290 nm, es decir, se trabaja con un proceso estocástico multivariado de $p = 4$ elementos por vector de observaciones, $\mathbf{Y} = (Y_1, Y_2, Y_3, Y_4)$, donde Y_j , $j = 1, 2, 3, 4$, será el espectro de emisión para una longitud de onda específica 230, 240, 255 y 290 nm. Estos datos están organizados en una matriz de tamaño 268×4 , donde cada fila tendrá los espectros de emisión de la i -ésima muestra, medida desde 275-560 nm, cuando la muestra ha sido excitada con luz en longitudes de onda de 230, 240, 255 y 290 nm. Cada columna de esta matriz contiene los perfiles correspondientes a una longitud de onda excitada (Y_j).

Cartas de Control		
$T_M^2 F$	fAO	MFHD
33.33857	6.402189	0.1460389

TABLA 4.1. Límites de control para las cartas $T_M^2 F$, fAO y MFHD, aplicadas al conjunto de datos de fluorescencia en el azúcar ajustados mediante suavizamiento spline

Dado que en este conjunto de datos no se tiene una clasificación entre datos dentro y fuera de control, se toman dos subconjuntos de datos D_1 y D_2 , de tamaños 200×4 y 68×4 , respectivamente. El primer conjunto, D_1 , es empleado para estimar los parámetros de las cartas en fase I, mientras que el conjunto D_2 es empleado para hacer monitoreo en fase II. Los datos se suavizan empleando 20 bases y polinomios de orden 6, bajo el método de suavizamiento spline.

En la gráfica 4.1 se presenta una muestra de 25 curvas del conjunto de datos suavizados en fase I (D_1), para longitudes de onda específica 230 y 290 nm. Una vez suavizados

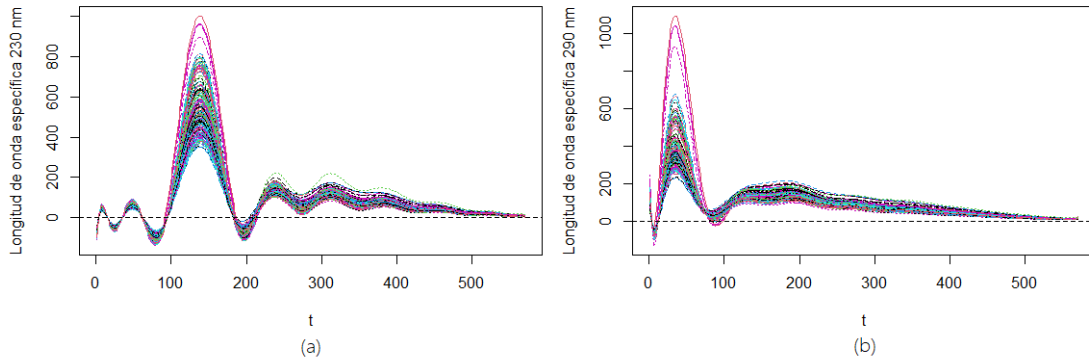


FIGURA 4.1. Muestra de 25 curvas estimadas para longitudes de onda 230 nm (a) y 290 nm (b) mediante suavizamiento spline, del conjunto de datos de fluorescencia en la producción de azúcar, en fase I

los datos, mediante suavizamiento spline, se procede a calcular los límites de control para las cartas $T_M^2 F$, fAO y MFHD, los cuales se presentan en la Tabla 4.1 Después de calcular los límites de control, se procede a monitorear los datos de D_2 de acuerdo a los límites calculados con base en D_1 . Las cartas de control asociadas a la fase II se presentan en el gráfico 4.2. En este caso se escogen $K = 4$ componentes principales, que retienen más del 99% de la varianza, y dado que entre $K = 4$ y $K = 5$ no hay más de 1% de ganancia en la retención de varianza. Como se puede observar en el gráfico 4.2, la carta de control $T_M^2 F$ identificó los perfiles multivariados funcionales 16 y 53 fuera de control en fase II, mientras que las cartas fAO y MFHD solamente identificaron la curva 53.

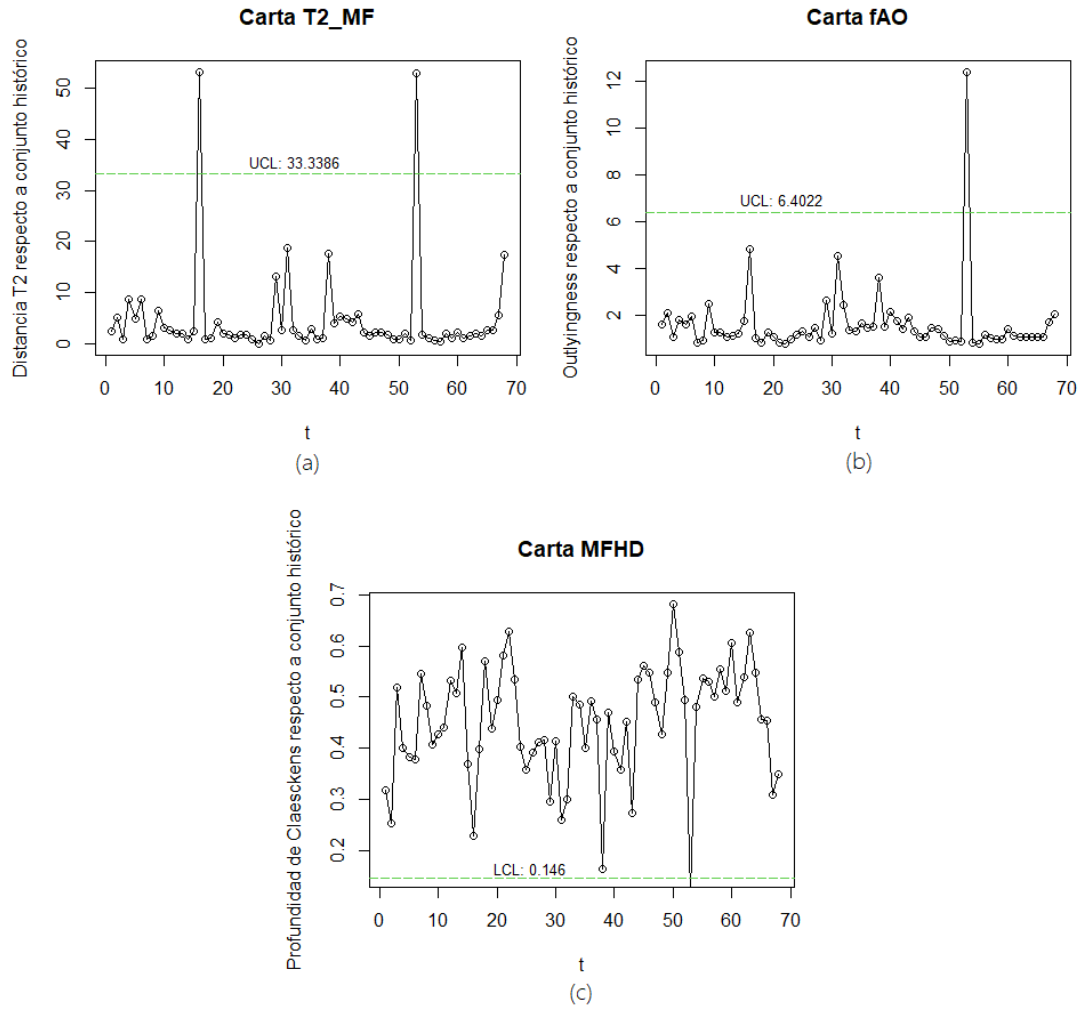


FIGURA 4.2. Cartas T^2_{MF} , fAO y MFHD en fase II, para el conjunto de datos D_2

Conclusiones

- Las cartas T_{MF}^2 y fAO presentaron potencias más altas que las demás para cambios de forma, cuando se monitorea un dato funcional univariado transformado en multivariado, empleando su primera y segunda derivada. Para el caso de cambios de magnitud se evidenciaron mejores resultados bajo las cartas fAO y F. Esto indica que empleando las derivadas de un proceso univariado se obtiene mayor información, lo cual permite identificar más rápido observaciones fuera de control, es decir, que resulta más potente derivar el proceso univariado y evaluar cartas de control multivariadas.
- Uno de los factores más importantes para evaluar la distancia funcional de Mahalanobis, univariada o multivariada, es la selección de las componentes a retener. Cuando se seleccionan muchos valores se presentan problemas en los eigenvalores, dado que estos son denominadores y aumentan las estadísticas de prueba. También se debe tener mucha precaución con la elección de las componentes a retener, dado que asintóticamente en algunos casos la distribución dependerá de este valor y esto influye en el cálculo de falsas alarmas en fase I y del ARL_0 en fase II.
- La carta multivariada más potente propuesta en esta Tesis fue la fAO, empleando la atipicidad ajustada funcional, seguida por la medida T^2 multivariada funcional. Estas dos medidas se calcularon sobre tamaños de muestras grandes en fase I, lo cual permitió tener valores de falsas alarmas cercanos al propuesto inicialmente con un valor de 0.005, que implica un ARL de 200. Esta condición indica que estas son cartas potentes con tamaños de muestra históricos grandes.
- En general la carta con menor potencia fue la que empleó la profundidad de Claeskens, lo cual puede verse influenciado por la selección del conjunto de pesos que se calculan para ponderar las observaciones. A su vez, esta estadística resultó ser una de las de mayor exigencia computacional.
- En general se observa que las cartas propuestas en esta tesis, a excepción de la profundidad de Claeskens, identificaron cambios pequeños en los procesos, comparadas con otras cartas ya existentes y aceptadas en la literatura.

Trabajo futuro

- Comparar las propuestas con otros métodos de monitoreo de perfiles propuestos en la literatura y optimizar computacionalmente las propuestas en estas tesis, dado que su costo ha sido bastante alto y esto ha impactado en el estudio de más factores y combinaciones.
- Identificar estadísticas para datos funcionales multivariados, o ajustar las existentes, para desarrollar cartas que empleen tamaños de muestra históricos pequeños.
- Evaluar el comportamiento de las cartas propuestas en esta tesis, cuando no se cumpla el supuesto de independencia entre las observaciones de los datos funcionales multivariados. Por ejemplo, series de tiempo funcionales o procesos espacio-temporales funcionales.
- Desarrollar cartas donde los procesos que componen el dato funcional multivariado no se encuentren en el mismo intervalo compacto, incluso que cambien de dimensión. Esto permitiría realizar cartas donde se tengan, por ejemplo, combinaciones entre curvas e imágenes, para refinar el control estadístico de procesos mediante ajuste de datos funcionales.

Bibliografía

- Atashgar, K. and Zargarabadi, O. Monitoring multivariate profile data in plastic parts manufacturing industries: An intelligently data processing. *Journal of Industrial Information Integration*, 8:38 – 48, 2017. ISSN 2452-414X. doi: <https://doi.org/10.1016/j.jii.2017.06.003>. URL <http://www.sciencedirect.com/science/article/pii/S2452414X16300942>.
- Berrendero, J. R., Justel, A., and Svarc, M. Principal components for multivariate functional data. *Computational Statistics & Data Analysis*, 55(9):2619–2634, 2011.
- Boone, J. and Chakraborti, S. Two simple shewhart-type multivariate nonparametric control charts. *Applied Stochastic Models in Business and Industry*, 28(2):130–140, 2012.
- Brys, G., Hubert, M., and Rousseeuw, P. J. A robustification of independent component analysis. *Journal of Chemometrics*, 19(5-7):364–375, 2005. doi: <https://doi.org/10.1002/cem.940>. URL <https://analyticalsciencejournals.onlinelibrary.wiley.com/doi/abs/10.1002/cem.940>.
- Brys, G., Hubert, M., and Struyf, A. A robust measure of skewness. *Journal of Computational and Graphical Statistics*, 13(4):996–1017, 2004.
- Chuang, S. C., Hung, Y. C., Tsai, W.-C., and Yang, S.-F. A framework for nonparametric profile monitoring. *Computers & Industrial Engineering*, 64(1):482 – 491, 2013. ISSN 0360-8352. doi: <https://doi.org/10.1016/j.cie.2012.08.006>. URL <http://www.sciencedirect.com/science/article/pii/S0360835212002057>.
- Claeskens, G., Hubert, M., Slaets, L., and Vakili, K. Multivariate functional half-space depth. *Journal of the American Statistical Association*, 109(505):411–423, 2014. doi: 10.1080/01621459.2013.856795. URL <https://doi.org/10.1080/01621459.2013.856795>.
- Crosier, R. B. Multivariate generalizations of cumulative sum quality-control schemes. *Technometrics*, 30(3):291–303, 1988.
- Cuevas, A., Febrero, M., and Fraiman, R. On the use of the bootstrap for estimating functions with functional data. *Computational Statistics & Data Analysis*, 51(2):1063 – 1074, 2006. ISSN 0167-9473. doi: <https://doi.org/10.1016/j.csda.2005.10.012>. URL <http://www.sciencedirect.com/science/article/pii/S0167947305002793>.

- Cuevas, A., Febrero, M., and Fraiman, R. Robust estimation and classification for functional data via projection-based depth notions. *Computational Statistics*, 22(3): 481–496, Sep 2007. ISSN 1613-9658. doi: 10.1007/s00180-007-0053-0. URL <https://doi.org/10.1007/s00180-007-0053-0>.
- Dai, W. and Genton, M. Directional outlyingness for multivariate functional data. *Computational Statistics and Data Analysis*, 131:50 – 65, 2019. ISSN 0167-9473. doi: <https://doi.org/10.1016/j.csda.2018.03.017>. URL <http://www.sciencedirect.com/science/article/pii/S016794731830077X>. High-dimensional and functional data analysis.
- Fassò, A., Toccu, M., and Magno, M. Functional control charts and health monitoring of steam sterilizers. *Quality and Reliability Engineering International*, 32(6):2081–2091, 2016.
- Ferraty, F. and Vieu, P. *Nonparametric Functional Data Analysis: Theory and Practice*. Springer Series in Statistics. Springer New York, 2010. ISBN 9781441921413. URL <https://books.google.com.co/books?id=S1VWcgAACAAJ>.
- Galeano, P., Joseph, E., and Lillo, R. E. The Mahalanobis distance for functional data with applications to classification. *Technometrics*, 57(2):281–291, 2015.
- Ghosh, M., Li, Y., Zeng, L., Zhang, Z., and Zhou, Q. Modeling multivariate profiles using gaussian process-controlled B-splines. *IISE Transactions*, 0(0):1–12, 2020. doi: 10.1080/24725854.2020.1798038. URL <https://doi.org/10.1080/24725854.2020.1798038>.
- Górecki, T., Krzyśko, M., Waszak, L., and Wołyński, W. Selected statistical methods of data analysis for multivariate functional data. *Statistical Papers*, pages 1–30, 2016.
- Grasso, M., Colosimo, B., and Pacella, M. Profile monitoring via sensor fusion: The use of PCA methods for multi-channel data. *International Journal of Production Research*, 02 2014. doi: 10.1080/00207543.2014.916431.
- Guevara, R. and Vargas, J. Evaluation of process capability in multivariate nonlinear profiles. *Journal of Statistical Computation and Simulation*, 86(12):2411–2428, 2016.
- Happ, C. and Greven, S. Multivariate functional principal component analysis for data observed on different (dimensional) domains. *Journal of the American Statistical Association*, 113(522):649–659, 2018. doi: 10.1080/01621459.2016.1273115. URL <https://doi.org/10.1080/01621459.2016.1273115>.
- Horváth, L. and Kokoszka, P. *Inference for Functional Data with Applications*. Springer Series in Statistics. Springer New York, 2012. ISBN 9781461436546. URL <https://books.google.com.co/books?id=PGOCTgAACAAJ>.
- Hubert, M. Data depth: Robust multivariate analysis, computational geometry and applications by Liu, R., Serfling, R., and Souvaine, D. L. *Biometrics*, 64(2):655–656, 2008. doi: 10.1111/j.1541-0420.2008.01026_6.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1541-0420.2008.01026_6.x.
- Hubert, M., Rousseeuw, P., and Segaeert, P. Multivariate functional outlier detection. *Statistical Methods & Applications*, 24:177–202, 2015.

- Hubert, M., Rousseeuw, P., and Segaert, P. Multivariate and functional classification using depth and distance. *Advances in Data Analysis and Classification*, 11(3):445–466, Sep 2017. ISSN 1862-5355. doi: 10.1007/s11634-016-0269-3. URL <https://doi.org/10.1007/s11634-016-0269-3>.
- Ieva, F. and Paganoni, A. M. Depth measures for multivariate functional data. *Communications in Statistics - Theory and Methods*, 42(7):1265–1276, 2013. doi: 10.1080/03610926.2012.746368. URL <https://doi.org/10.1080/03610926.2012.746368>.
- Jacques, J. and Preda, C. Model-based clustering for multivariate functional data. *Computational Statistics & Data Analysis*, 71:92–106, 2014.
- Jahani, S., Kontar, R., Veeramani, D., and Zhou, S. Statistical monitoring of multiple profiles simultaneously using gaussian processes. *Quality and Reliability Engineering International*, 34(8):1510–1529, 2018. doi: 10.1002/qre.2326. URL <https://onlinelibrary.wiley.com/doi/abs/10.1002/qre.2326>.
- Jones-Farmer, L. A., Woodall, W. H., Steiner, S. H., and Champ, C. W. An overview of Phase I analysis for process improvement and monitoring. *Journal of Quality Technology*, 46(3):265–280, 2014. doi: 10.1080/00224065.2014.11917969. URL <https://doi.org/10.1080/00224065.2014.11917969>.
- Kang, L. and Albin, S. L. On-line monitoring when the process yields a linear profile. *Journal of Quality Technology*, 32(4):418, 2000.
- Kokoszka, P. and Reimherr, M. *Introduction to Functional Data Analysis*. Chapman & Hall / CRC numerical analysis and scientific computing. CRC Press, 2017. ISBN 9781498746342. URL <https://books.google.com/books?id=HIXIvgAACAAJ>.
- Li, Z., Dai, Y., and Wang, Z. Multivariate change point control chart based on data depth for Phase I analysis. *Communications in Statistics - Simulation and Computation*, 43(6):1490–1507, 2014. doi: 10.1080/03610918.2012.735319. URL <https://doi.org/10.1080/03610918.2012.735319>.
- Liu, R. Y., Parelius, J. M., and Singh, K. Multivariate analysis by data depth: descriptive statistics, graphics and inference, (with discussion and a rejoinder by liu and singh). *Ann. Statist.*, 27(3):783–858, 06 1999. doi: 10.1214/aos/1018031260. URL <https://doi.org/10.1214/aos/1018031260>.
- López-Pintado, S. and Romo, J. On the concept of depth for functional data. *Journal of the American Statistical Association*, 104(486):718–734, 2009. doi: 10.1198/jasa.2009.0108. URL <https://doi.org/10.1198/jasa.2009.0108>.
- Lowry, C. A., Woodall, W. H., Champ, C. W., and Rigdon, S. E. A multivariate exponentially weighted moving average control chart. *Technometrics*, 34(1):46–53, 1992.
- Maleki, M., Amiri, A., and Castagliola, P. An overview on recent profile monitoring papers (2008-2018) based on conceptual classification scheme. *Computers & Industrial Engineering*, 126:705 – 728, 2018. ISSN 0360-8352.
- Montgomery, D. C. *Introduction to statistical quality control*. John Wiley & Sons, 2007.

- Munck, L., Norgaard, L., Engelsen, S., Bro, R., and Andersson, C. Chemometrics in food science—a demonstration of the feasibility of a highly exploratory, inductive evaluation strategy of fundamental scientific significance. *Chemometrics and Intelligent Laboratory Systems*, 44(1):31–60, 1998. ISSN 0169-7439. doi: [https://doi.org/10.1016/S0169-7439\(98\)00074-4](https://doi.org/10.1016/S0169-7439(98)00074-4). URL <https://www.sciencedirect.com/science/article/pii/S0169743998000744>.
- Noorossana, R., Saghaei, A., and Amiri, A. *Statistical analysis of profile monitoring*, volume 865. John Wiley & Sons, 2011.
- Pan, J.-N., Li, C.-I., and Lu, M. Z. Detecting the process changes for multivariate nonlinear profile data. *Quality and Reliability Engineering International*, 35(6):1890–1910, 2019. doi: 10.1002/qre.2482. URL <https://onlinelibrary.wiley.com/doi/abs/10.1002/qre.2482>.
- Paynabar, K., Zou, C., and Qiu, P. A change-point approach for Phase-I analysis in multivariate profile monitoring and diagnosis. *Technometrics*, 58(2):191–204, 2016. doi: 10.1080/00401706.2015.1042168. URL <https://doi.org/10.1080/00401706.2015.1042168>.
- Peña, D. *Análisis multivariante de datos*. McGraw-Hill Interamericana de España S.L., 2002. ISBN 9788448136109.
- Pignatiello, J. J. and Runger, G. C. Comparisons of multivariate cusum charts. *Journal of quality technology*, 22(3):173–186, 1990.
- Qiu, P. *Introduction to statistical process control*. CRC Press, 2013.
- Qiu, P., Zou, C., and Wang, Z. Nonparametric profile monitoring by mixed effects modeling. *Technometrics*, 52(3):265–277, 2010.
- R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2016. URL <https://www.R-project.org/>.
- Ramsay, J. and Silverman, B. *Functional data analysis*. Springer Series in Statistics. Springer, 2ed edition, 2005. ISBN 038740080X,9780387400808.
- Schabenberger, O. and Pierce, F. J. *Contemporary statistical models for the plant and soil sciences*. CRC press, 2001.
- Sheu, S. H., Ouyoung, C. W., and Hsu, T. S. Phase II statistical process control for functional data. *Journal of Statistical Computation and Simulation*, 83(11):2144–2159, 2013.
- Shiau, J.-J. H., Huang, H.-L., Lin, S.-H., and Tsai, M.-Y. Monitoring nonlinear profiles with random effects by nonparametric regression. *Communications in Statistics Theory and Methods*, 38(10):1664–1679, 2009.
- Tarabelloni, N., Biasi, R., Paganoni, A., and Ieva, F. Multivariate functional data depth measure based on variance-covariance operators. 07 2014.
- Wang, K. and Tsung, F. Hierarchical sparse functional principal component analysis for multistage multivariate profile data. *IISE Transactions*, 0(0):1–16, 2020. doi: 10.1080/24725854.2020.1738599. URL <https://doi.org/10.1080/24725854.2020.1738599>.

-
- Wang, Y., Mei, Y., and Paynabar, K. Thresholded multivariate principal component analysis for Phase I multichannel profile monitoring. *Technometrics*, 0(0):1–13, 2018. doi: 10.1080/00401706.2017.1375993. URL <https://doi.org/10.1080/00401706.2017.1375993>.
- Williams, J. D., Woodall, W. H., and Birch, J. B. Statistical monitoring of nonlinear product and process quality profiles. *Quality and Reliability Engineering International*, 23(8):925–941, 2007.
- Yeh, A. B., Huwang, L., and Li, Y.-M. Profile monitoring for a binary response. *IIE Transactions*, 41(11):931–941, 2009.
- Zhang, C., Yan, H., Lee, S., and Shi, J. Weakly correlated profile monitoring based on sparse multi-channel functional principal component analysis. *IIE Transactions*, 50(10):878–891, 2018. doi: 10.1080/24725854.2018.1451012. URL <https://doi.org/10.1080/24725854.2018.1451012>.
- Zhang, J., Ren, H., Yao, R., Zou, C., and Wang, Z. Phase I analysis of multivariate profiles based on regression adjustment. *Comput. Ind. Eng.*, 85(C):132–144, July 2015. ISSN 0360-8352. doi: 10.1016/j.cie.2015.02.025. URL <http://dx.doi.org/10.1016/j.cie.2015.02.025>.
- Zou, C., Tsung, F., and Wang, Z. Monitoring profiles based on nonparametric regression methods. *Technometrics*, 50(4):512–526, 2008. doi: 10.1198/004017008000000433. URL <https://doi.org/10.1198/004017008000000433>.
- Zou, C., Wang, Z., and Tsung, F. A spatial rank-based multivariate EWMA control chart. *Naval Research Logistics (NRL)*, 59(2):91–110, 2012.