

RNPS No. 2142 ISSN 2072-6287 Versión Digital

# REPORTE TÉCNICO Reconocimiento de Patrones

## Nuevos métodos para la calibración estadística basada en datos univariados, multivariados y funcionales

Noslen Hernández González

RT\_046

enero 2012

7ma. No. 21812 e/218 y 222, Rpto. Siboney, Playa; La Habana. Cuba. C.P. 12200 www.cenatav.co.cu CENATAV CENATAV Centro de Aplicaciones de Tecnologías de Avanzada MINISTERIO DE LAINDUSTRIARÁSICA RNPS No. 2142 ISSN 2072-6287 Versión Digital

# REPORTE TÉCNICO Reconocimiento de Patrones

### Nuevos métodos para la calibración estadística basada en datos univariados, multivariados y funcionales

Noslen Hernández González

RT\_046

enero 2012

7ma. No. 21812 e/218 y 222, Rpto. Siboney, Playa; La Habana. Cuba. C.P. 12200 www.cenatav.co.cu Centro de Aplicaciones de Tecnologías de Avanzada Departamento de Reconocimiento de Patrones



# Nuevos métodos para la calibración estadística basada en datos univariados, multivariados y funcionales

Tesis presentada en opción al grado científico de Doctor en Ciencias Matemáticas

por

## Ing. Noslen HERNÁNDEZ GONZÁLEZ

Supervisada por : Dr. Rolando J. BISCAY LIRIO Dra. Isneri TALAVERA BUSTAMANTE

Instituto de Cibernética, Matemática y Física, ICIMAF Ciudad de La Habana 2010

A mi mamá y a mi papá, los pilares de esta travesía

Un camino de alegrías, de sorpresas, de ilusiones, de tropiezos, de imponentes retos, de increíbles experiencias, de incertidumbre, de placer... un camino que comienza cuando crees que se acaba, un camino que resulta intransitable si se recorre solo... a todas las personas maravillosas que han llegado para iluminar este camino, van mis AGRADECIMIENTOS.

Gracias mami, papi y Yoya, por existir para mí, por su amor y su guía, por ser mi tridente mágico con que desvanezco toda dificultad, todo obstáculo, toda pena... por hacerme tan feliz.

Gracias Isne (¡la tutora!) por iniciarme en el mundo de la Quimiometría, por descubrir nuevos problemas a resolver, nuevos retos. Por tu entusiasmo, tu apoyo DIARIO y tu confianza. Por la ilusión de mis primeros pasos como profesional: mi primer paper, mi primer seminario, mi primer proyecto.

Gracias Biscay (jel tutor!) por hacer realidad mi sueño de regresar al sendero de las Matemáticas, por tu sabiduría infinita, por tu "me explico" en lugar de un "entendiste", por tu paciencia. Por abrirme las puertas del universo de todo lo que queda por aprender. Por la magia con que aligeras la dureza del trabajo, porque a tu lado el trabajo es placer.

Gracias Dina, por mostrarme el arte de hacer ciencia y divertirse, de crear con amor, por enseñarme que las ideas más simples pueden esconder soluciones geniales.

Merci, Nathalie, pour l'ocassion de travailler ensemble, d'apprendre des choses nouvelles et intéressantes. Merci pour ton hospitalité pendant mes séjours à Toulousse. Et sourtout, en plus de ton professionnalisme, merci de me donner ton amitié.

Agradeço a professora Márcia pelos ensinamentos no curso de Quimiometria, conhecimentos, experiências e hospitalidade que recebi durante minha estadia no LQTA-Campinas-Brasil. Agradeço também a todos aqueles que tornaram meus dias felices e agradáveis no Brasil: Cel, Flávia, Wesley, Kerly, Euzebín, Daniel, Tahiana, Eduardo, Teófilo, Rudolf, JP, Rodrigo.

Gracias a todos mis BUENOS amigos, que en esta ocasión también hicieron de las suyas:

A Heydi por aguantar todas mis malcriadeces en etapas de STRESS, por escuchar, por pelear, por la dicha de ir brincando juntas los mismos muros... y siempre recibir su empujoncito. A Jyrko por siempre hacerme creer que podía hacer matemáticas.

A Candy por ponerle a la tesis su toque literario.

A Annette por su faceta de diseñadora y su francés.

A Diana y Oneisys por su ayuda incondicional, por los muchos fuegos que han apagado por mí.

A Irén por la compañía de siempre.

Gracias a mis compañeros de Departamento (físico, lógico y guaroso) Garea, Gilito, Mabel, Gabriel, Calvo, Javier, Chang, Yeni, el cuadro, Sandro, Richín y LA FIGURA por todo su apoyo. A Maité, por esperar con paciencia siempre el último minuto; a Adán por su disposición. A mis colegas de ICT. A todos los que con cariño, me han brindado su sostén alguna vez.

Gracias a mi equipo de proyecto Bustio, Dania, Yoanna y Alain por la comprensión y el esfuerzo adicional que han asumido para ayudarme.

Gracias a los profesores Minerva Montero y Li-Vang Lozada por contribuir al mejoramiento de esta tesis con sus oponencias. A muchas personas del ICIMAF por su granito de arena. A Martica, por sus consejos, ¡claves! y su alegría.

Gracias a mis super-abuelos Rita, Gladys, Arturo y Juan por velar por todos mis pasos, por no escatimar esfuerzos, por la satisfacción de ver, al escucharlos hablar sobre mis proyectos, aspiraciones y tareas, que su preocupación y su ocupación rompen las barreras del desconocimiento.

Gracias a mi aguerrida y numerosa familia, la genética y la social, siempre en pie de combate, siempre en mesa redonda: Yose, Yanet, Yadira, Pimpo, mis tíos, Wimbi, mi tía Nelsa, Cary, la familia Bombalier, Annalis, Ela, Marta, la Mily, Ernesto, Deysi, la retaguardia francesa Salomón, Li-Vang, Ivonne, Alien.

Le agradezco a mi país el lugar cimero en que tiene la investigación científica, le agradezco que existan proyectos como CENATAV, con grandes potencialidades para la formación profesional.

# SÍNTESIS

En esta tesis se investiga el problema de la calibración estadística, motivado por la solución de problemas de calibración con datos provenientes del análisis químico instrumental como es el caso de los datos espectroscópicos, cromatográficos, entre otros. Se desarrollan nuevos métodos de calibración para diferentes espacios muestrales: univariados, multivariados y funcionales. En particular, para la calibración univariada lineal, se propone un enfoque predictivo no Bayesiano que mitiga algunas de las limitaciones de los enfoques frecuentistas existentes, como son: la obtención únicamente de estimaciones puntuales y estimadores con propiedades estadísticas indeseables o virtudes sólo fundamentadas asintóticamente. También se estudian algunas propiedades del estimador derivado. Para datos multivariados, se introduce el uso de las Máquinas de Vectores Relevantes para Regresión en la solución de problemas de calibración. Por último, para la calibración con datos funcionales, se propone un método de regresión no paramétrico ralo que consiste en la extensión de las Máquinas de Vectores Soportes a datos funcionales, junto a una nueva clase de funciones núcleos, apropiada para este tipo de dato. Además, se desarrolla un nuevo método no paramétrico de regresión inversa basado en la densidad, que ofrece ventajas con respecto a los métodos de regresión inversa funcional y otros no paramétricos reportados en la literatura. El comportamiento de todos los enfoques propuestos se ilustra a través de estudios de simulación o la aplicación a datos reales.

# ÍNDICE

LI	LISTADO DE SIGLAS 1						
INTRODUCCIÓN 3							
1.	UN	UN ENFOQUE PREDICTIVO NO BAYESIANO PARA LA CA-					
	LIB	RACIÓN UNIVARIADA	13				
	1.1.	Revisión de los enfoques previos	13				
	1.2.	Estimador predictivo no Bayesiano	18				
	1.3.	Propiedades del estimador predictivo no Bayesiano	21				
		1.3.1. Desarrollo asintótico del estimador	21				
		1.3.2. Desarrollo asintótico del error cuadrático medio del estimador	23				
		1.3.3. Finitud de los momentos del estimador	26				
		1.3.4. Consistencia del estimador	28				
	1.4.	Estudio de simulación	30				
2.	MÁ	MÁQUINAS DE VECTORES RELEVANTES PARA LA CALI-					
	BR.	ACIÓN MULTIVARIADA	39				
	2.1.	Introducción	39				
	2.2.	Máquinas de vectores relevantes para regresión	41				
	2.3.	Aplicación a datos reales	43				
		2.3.1. Conjunto de datos "Wülfert"	44				
		2.3.2. Conjunto de datos "Tecator"	48				
		2.3.3. Conjunto de datos "Tablet"	51				
3.	MÁ	MÁQUINAS DE VECTORES SOPORTES PARA DATOS FUN-					
	CIC	CIONALES 55					
	3.1.	Revisión de los enfoques previos de regresión funcional	55				

	3.2.	.2. Conceptos preliminares sobre la regresión no paramétrica por regula-							
		rización en espacios de Hilbert con núcleos reproductores							
	3.3. Máquinas de vectores soportes funcionales								
		3.3.1.	Definición del método FSVR	61					
		3.3.2.	Núcleos definidos no negativos para espacios funcionales	63					
	3.4.	Result	ados en el conjunto de datos "Tecator"	65					
4.	REC	EGRESIÓN INVERSA BASADA EN LA DENSIDAD PARA DA-							
	TOS	5 FUN	CIONALES	69					
	4.1.	Introd	ucción	69					
	4.2.	Regres	sión inversa basada en la densidad	70					
		4.2.1.	Definición de DBIR en un marco abstracto	70					
		4.2.2.	Especificación para el caso Gaussiano	72					
	4.3. Consistencia del estimador		$tencia del estimador \dots \dots$	73					
		4.3.1.	Notaciones	73					
		4.3.2.	Parte 1: Convergencia de $\hat{r}_{\eta_n}(x)$ y razón de convergencia 	74					
		4.3.3.	Parte 2: Convergencia de $\hat{\Gamma}$	76					
		4.3.4.	Parte 3: Convergencia de $\hat{f}(y \swarrow x)$	81					
		4.3.5.	Conclusión: Consistencia de $\hat{\gamma}(y)$	86					
	4.4.	4.4. Estudio de simulación		88					
		4.4.1.	Generación de los datos $\ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots$	88					
		4.4.2.	Resultados de la simulación	90					
	4.5.	Result	ados en el conjunto de datos "Tecator"	94					
CONCLUSIONES Y RECOMENDACIONES 97									
BIBLIOGRAFÍA 1									
А.	A. Términos químicos								
в.	$\mathbf{Asp}$	ectos	computacionales	113					

# Listado de siglas

$\mathbf{CR}$	Regresión Continua
DBIR	Regresión Inversa Basada en la Densidad
FDA	Análisis de Datos Funcionales
FIR	Regresión Inversa Funcional
FSVR	Máquinas de Vectores Soportes Funcionales
KL	Kullback-Leibler
KPLS	Mínimos Cuadrados Parciales con Núcleos
LSSVM	Máquinas de Vectores Soportes de Mínimos Cuadrados
MSC	Corrección Multiplicativa de la Señal
MSE	Error Cuadrático Medio
NN	Redes Neuronales
NSR	Relación Ruido-Señal
NWK	Núcleo de Nadaraya-Watson
PCR	Regresión por Componentes Principales
PLS	Mínimos Cuadrados Parciales
RBF	Funciones de Bases Radiales
RKHS	Espacios de Hilbert con Núcleos Reproductores
RMSE	Raíz del Error Cuadrático Medio
RMSECV	Raíz del Error Cuadrático Medio de Validación Cruzada
RMSEP	Raíz del Error Cuadrático Medio de Predicción
RR	Regresión Ridge
RVMR	Máquinas de Vectores Relevantes para Regresión
SIR	Regresión Inversa Partida
SVR	Máquina de Vectores Soportes

# INTRODUCCIÓN

La calibración estadística ha sido estudiada durante décadas [69] y encuentra múltiples aplicaciones en áreas como la instrumentación, la farmacología y la quimiometría [11, 61, 63, 103]. El problema de la calibración estadística puede describirse de la siguiente forma: una variable aleatoria observable Y está relacionada con una variable de interés X de acuerdo con un modelo estadístico especificado por una función de densidad de probabilidad condicional  $f(\cdot; X, \boldsymbol{\theta})$ , la cual depende además de un parámetro desconocido  $\boldsymbol{\theta}$ . Se dispone de una muestra  $\mathcal{D}$  de observaciones independientes  $(x_1, y_1), ..., (x_n, y_n)$  de (X, Y), conocida como muestra de calibración. Dada una observación nueva (o futura)  $y_0$  de Y que corresponde a un valor desconocido  $x_0$  de X, el problema de calibración consiste en hacer inferencia estadística sobre  $x_0$  sobre la base del modelo estadístico dado, de los datos previos  $\mathcal{D}$  y de la nueva observación  $y_0$ .

La motivación de esta tesis lo constituye la solución de problemas de calibración con datos provenientes del análisis químico instrumental (*e.g.*, espectroscópicos, cromatográficos, entre otros) [63]. Los químicos usualmente desean estudiar la composición química X de una mezcla (*i.e.*, una combinación de concentraciones de compuestos) a partir de mediciones espectrales Y realizadas a la misma (*e.g.*, espectro de absorbancia de la luz a diferentes longitudes de onda). En este contexto, la densidad condicional  $f(\cdot; X, \theta)$  representa el proceso físico de generación de los datos en el cual el espectro Y medido por el instrumento está determinado por las concentraciones X de los compuestos presentes en la mezcla. El problema de calibración consiste entonces en: dada una observación de un espectro  $y_0$  correspondiente a una mezcla nueva, estimar su composición química  $x_0$  sobre la base de observaciones (pasadas) de composiciones químicas y espectros ( $x_i, y_i$ ), i = 1, ...n, en la muestra de calibración.

En lo adelante, se restringe la atención al caso en que la variable de interés X en la calibración es una variable con valores en los reales (*e.g.*, la concentración de un solo compuesto a determinar). Pero aún con esta simplificación, desde el punto de vista

estadístico, la problemática de la calibración es notablemente variada y multifacética. Esto se debe principalmente a los siguientes aspectos:

- 1. En general, la observación Y puede tomar valores en distintos tipos de espacios muestrales  $\mathcal{Y}$ . Específicamente, puede ser un dato univariado  $Y \in \mathbb{R}$  (e.g., valor de absorbancia a una frecuencia específica del espectro), un dato multivariado  $Y \in \mathbb{R}^d$  (e.g., vector de valores de absorbancia en múltiples frecuencias) o un dato funcional  $Y \in L_2([a, b])$  (e.g., función de absorbancia en un intervalo de frecuencias de interés).
- Los valores  $x_i$  de la variable X en la muestra de calibración se pueden fijar en 2.el diseño por el investigador (situación conocida como calibración controlada) o pueden constituir una muestra aleatoria obtenida bajo determinadas condiciones (situación llamada *calibración aleatoria o natural*) [69]. En el caso de la calibración natural se sobrentiende que los valores futuros  $X_0$  de la variable X tienen la misma distribución f(x) que los valores  $x_i$  en la muestra de calibración, por lo cual resulta apropiado considerar un criterio de riesgo Bayesiano  $R(\widehat{X}_0) = \mathbb{E}l(X_0, \widehat{X}_0)$  para evaluar la calidad de una propuesta de estimador  $\widehat{X}_0$ , según una función de pérdida l especificada por el investigador; e.q., el riesgo Bayesiano cuadrático  $R(\hat{X}_0) = \mathbb{E}(X_0 - \hat{X}_0)^2$ . Nótese que aquí el valor esperado se toma no sólo con respecto a la distribución condicional de la observación  $Y_0$ según el modelo  $f(\cdot; X_0, \boldsymbol{\theta})$ , sino también con respecto a la distribución f(x)de  $X_0$ . Por el contrario, en el caso de la calibración controlada, los valores futuros  $X_0$  de la variable X no necesariamente siguen la misma distribución que en la muestra de calibración, por tanto, un criterio de riesgo no Bayesiano es más adecuado. En dependencia de la situación práctica, se utilizan diversos criterios Bayesianos y no Bayesianos para evaluar los métodos estadísticos de calibración.
- 3. La dependencia entre las variables  $X \neq Y$  puede ser *lineal* o no lineal.

Ante esta variada problemática se han elaborado diversos enfoques. A continuación se hará una revisión de los principales de ellos, y se destacarán los problemas de investigación abiertos que constituyen motivaciones de los objetivos de este trabajo de tesis.

Para el *problema de la calibración univariada lineal* los enfoques propuestos pueden agruparse en tres clases fundamentales: El método de máxima verosimilitud para estimar los parámetros  $\boldsymbol{\theta}$  y  $x_0$  conduce al llamado estimador clásico  $\hat{x}_C$  de  $x_0$  introducido por Eisenhart [24]. Este estimador presenta algunas propiedades favorables como son: ser consistente y que su error cuadrático medio asintótico es constante en todo el rango de posibles valores de  $x_0$ [4]. Sin embargo, se ha criticado grandemente debido a que no tiene momentos finitos y por tanto, su error cuadrático medio (MSE, del inglés <u>Mean Square Error</u>) es infinito [104].

#### b) Estimadores Bayesianos

Como alternativas al estimador clásico, se han propuesto varios estimadores que pueden justificarse desde una perspectiva Bayesiana. Uno de ellos es el estimador inverso  $\hat{x}_I$  introducido por Krutchkoff [52, 53] que se obtiene como un estimador Bayesiano bajo distribuciones a priori no informativas [1] y, alternativamente, minimizando (sobre cierta clase de estimadores lineales) versiones empíricas del riesgo Bayesiano, como son el criterio de validación cruzada [71] o el error compuesto [60].

Varios estimadores se reducen o son muy similares al estimador inverso bajo ciertas condiciones. Este es el caso de los estimadores propuestos por Halperin [36] y Hagwood [35]. Algunos estudios teóricos y de simulaciones publicados comparan el estimador clásico y el inverso teniendo en cuenta varios criterios [4, 16, 17, 51, 52, 53, 60, 84, 93, 99, 104]. De estas comparaciones se concluye que el estimador clásico, en comparación con el inverso, muestra mejor comportamiento para la extrapolación y peor comportamiento para la interpolación; el estimador  $\hat{x}_C$  es consistente, mientras que  $\hat{x}_I$  no lo es; a diferencia de  $\hat{x}_C$ , los momentos de  $\hat{x}_I$  y por tanto su MSE son finitos; además, en situaciones en que la razón señal-ruido es pequeña, el comportamiento de  $\hat{x}_C$  se deteriora de manera más drástica que el de  $\hat{x}_I$ .

#### c) Estimadores basados en correcciones asintóticas del estimador clásico

En la literatura se recogen correcciones o modificaciones al estimador clásico. Algunas de ellas consisten en sumas ponderadas de estimadores, como es el caso de los estimadores de Ali y Singh [2] y de Srivastava y Singh [87]. Otro caso es el estimador de Naszódi [68], el cual garantiza insesgadez asintótica hasta el orden 1/n. De lo anterior se deriva que, hasta el momento, los enfoques no Bayesianos existentes para el problema de la calibración univariada lineal presentan varias limitaciones:

- Poseen propiedades estadísticas marcadamente indeseables (como el estimador clásico) o tienen virtudes sólo fundamentadas por consideraciones asintóticas, no válidas para muestras finitas.
- II) Sólo brindan estimaciones puntuales y en ocasiones por intervalos, pero no otras formas de inferencia estadística como las basadas en la exploración de la forma de la función de verosimilitud.

Por tanto, la elaboración de un enfoque no Bayesiano (frecuentista) para la calibración univariada lineal que supere estas deficiencias constituye un problema abierto.

Para el problema de la calibración multivariada (i.e., cuando se tiene dato vectorial  $Y \in \mathbb{R}^d$ ) se ha propuesto una gran variedad de métodos. Estos, en su mayoría, están orientados a minimizar el riesgo Bayesiano y pueden agruparse en tres grandes clases:

a) Métodos de calibración multivariada lineales.

El estimador clásico de Eisenhart [24] se generaliza al caso multivariado [10], el cual puede obtenerse además, siguiendo el enfoque de verosimilitud bajo el supuesto de normalidad. Se ha propuesto otra variedad de estimadores lineales de  $x_0$  que intenta aproximar el valor esperado condicional  $\mathbb{E}(X \not/ y_0)$ , como es el caso de la generalización multivariada del estimador inverso [10]. Una dificultad que enfrentan los predictores lineales es la alta dimensión y la colinealidad de las variables debido a la alta dimensión en Y, especialmente en las aplicaciones en espectrometría. Varios métodos se han desarrollado para mitigar esta dificultad. Entre lo más usados se encuentran: la Regresión Ridge (RR, del inglés <u>Ridge Regression</u>) [49, 50], la Regresión por Componentes Principales (PCR, del inglés <u>Principal Component Regression</u>) [64], los Mínimos Cuadrados Parciales (PLS, del inglés <u>Partial Least Squares</u>) [39, 40, 108, 106, 107], y la Regresión Continua (CR, del inglés <u>Continuum Regression</u>) [32, 38, 88, 89].

b) Métodos de calibración multivariada no paramétricos.

Una desventaja de los predictores lineales es que no siempre pueden modelar las no linealidades existentes entre las variables  $X ext{ y } Y$ . Motivado por esto, se han propuesto varios métodos no paramétricos para el problema de calibración. Entre ellos se encuentran las Redes Neuronales (NN, del inglés <u>Neural Networks</u>) [6], los estimadores de Núcleos de Nadaraya-Watson (NWK, del inglés <u>Nadaraya-Watson Kernel</u>) [67], las Funciones de Bases Radiales (del inglés, <u>Radial Basis Functions</u>) [102], los Mínimos Cuadrados Parciales con Núcleo (KPLS, del inglés <u>Kernel Partial Least Squares</u>) [78], las Máquinas de Vectores Soportes (SVR, del inglés <u>Support Vector Regression</u>) [3, 19, 82, 85, 94, 100] y las Máquinas de Vectores Soportes de Mínimos Cuadrados (LSSVM, del inglés Least Square Support Vector Machines) [95, 90], entre otros.

#### c) Métodos de regresión inversa

Una metodología semi-paramétrica para la calibración multivariada denominada Regresión Inversa Partida (SIR, del inglés <u>Sliced Inverse Regression</u>) fue introducida por Li [57]. SIR es un método de reducción de dimensión que supone que toda la información en Y acerca de X puede obtenerse a través de su proyección en un subespacio de menor dimensión. La estimación del modelo se lleva a cabo mediante la estimación de la función de regresión inversa  $\mathbb{E}(Y \nearrow X)$  la cual, bajo ciertos supuestos, subyace en dicho subespacio. Una aplicación del método SIR en la solución de un problema de calibración en la industria del petróleo puede encontrarse en el trabajo publicado por Hernández y col. [48].

A pesar de contar con todos estos enfoques no paramétricos multivariados, aún no se ha explorado la utilidad del método Máquinas de Vectores Relevantes para Regresión (RVMR, del inglés <u>Relevance Vector Machines Regression</u>) [97] en la solución de problemas de calibración. El método RVMR resulta promisorio en este contexto porque ofrece una solución Bayesiana, permite el uso de funciones bases arbitrarias, no necesita ajustar muchos hiperparámetros (*i.e.*, parámetros de control especificados por el investigador) y conduce a soluciones ralas (ó <u>sparse</u> en inglés; *i.e.*, estimaciones dependientes de un gran número de parámetros pero muchos de ellos resultan tener valor cero). Por tanto, un problema abierto de interés es evaluar las posibles bondades del enfoque RVMR en la solución de problemas de calibración multivariada con datos reales.

Otro aspecto interesante a considerar es que como consecuencia del desarrollo vertiginoso de los instrumentos de medición, los cuales proveen una gran cantidad de datos como funciones digitalizadas de alta resolución, la observación Y suele representar una discretización de una función, lo cual no es considerado por los enfoques multivariados. Esto ha provocado que el Análisis de Datos Funcionales (FDA, del inglés Functional Data Analysis) [27, 76] se haya convertido en un campo de cre-

ciente investigación para la solución de problemas de calibración. En el contexto del FDA, cada dato individual se trata como una única entidad, descrita por una función continua que toma valores en los números reales, en lugar de por un vector de dimensión finita. Por tanto, los datos funcionales se consideran elementos en espacios de dimensión infinita, típicamente espacios de Hilbert  $\mathcal{Y}$  (e.g.,  $\mathcal{Y} = L_2[a, b]$ ).

Hasta este momento, los métodos empleados para el problema de calibración con datos funcionales han estado orientados a aproximar la función de regresión  $\gamma(y) = \mathbb{E}(X / Y = y)$ , *i.e.* a minimizar el riesgo Bayesiano cuadrático. Estos enfoques pueden ser clasificados en tres familias fundamentales:

#### a) Métodos de calibración funcionales lineales

Los primeros trabajos sobre calibración con datos funcionales estuvieron enfocados a modelos de regresión donde la función de regresión  $\gamma$  tiene forma lineal. Una revisión de los enfoques para estimar este modelo puede encontrarse en [76]; ver también [12, 13, 15, 18, 38, 62, 74].

#### b) Métodos de calibración funcionales no paramétricos

Una desventaja de los métodos lineales es que no pueden tratar las dependencias no lineales entre el predictor y la variable respuesta. Para superar esto, se han propuesto varios métodos no paramétricos.

Un primer enfoque, introducido por Ferraty y Vieu [27], es el uso de estimadores de núcleo para la regresión funcional. Este tipo de estimadores ofrece una gran flexibilidad para ajustar modelos no lineales. Sin embargo, sus propiedades de consistencia se han demostrado sólo para clases restringidas de núcleos, como son los de Tipos I y II, definidos en [27]. Además, la selección del valor del parámetro de ancho o escala del núcleo, constituye un problema difícil en el contexto del FDA, que aún es tema de investigación.

Otra clase de estimadores de regresión no paramétricos son las redes neuronales funcionales propuestas por Rossi y col. [79]. Este enfoque tiene la propiedad de aproximación universal, pero tiene el inconveniente de que depende de un gran número de parámetros cuya estimación conduce a problemas de múltiples mínimos locales. Además, debe ajustarse el número de neuronas (arquitectura de la red), lo cual es una tarea computacionalmente difícil. En esta misma dirección, se introdujeron versiones funcionales de las redes neuronales con bases radiales [80].

El marco general de aproximación de funciones en Espacios de Hilbert con Núcleos Reproductores (RKHS, del inglés Reproducing Kernel Hilbert Spaces) también se ha usado [73] para introducir estimadores de regresión para datos funcionales, en particular, extensiones a datos funcionales de las funciones de bases radiales.

Una ventaja importante del enfoque de aproximación de funciones en RKHS es que el estimador resultante es lineal en los parámetros a estimar; por tanto, su cálculo se reduce a resolver un problema algebraico lineal. Sin embargo, a diferencia de los métodos estándares para aproximar funciones multivariadas en RKHS (*e.g.*, splines multivariados y funciones de bases radiales multivariadas como los <u>thin-plate splines</u>), en el contexto del FDA, aún no se han definido las propiedades de suavidad del estimador de la función de regresión como un funcional en RKHS. Por tanto, la selección de núcleos y funciones de bases radiales adecuados según los criterios de suavidad es todavía un problema abierto.

Otro método de regresión funcional no paramétrico introducido recientemente es la regresión usando k-vecinos más cercanos. Tanto su consistencia, como su razón de convergencia se han estudiado [5, 56].

#### c) Métodos de Regresión Inversa Funcional

Recientemente, se introdujo una metodología alternativa que puede interpretarse como un compromiso entre los métodos paramétricos restrictivos (*e.g.*, regresión lineal) y los no paramétricos (*e.g.*, basados en núcleos). Este enfoque se conoce como Regresión Inversa Funcional (FIR, del inglés <u>Functional Inverse Regression</u>) [21, 29, 30], y constituye la versión funcional del método SIR, propuesto para modelos de regresión multivariados [57]. FIR supone que toda la información en Y acerca de X puede obtenerse a través de su proyección en un subespacio de dimensión finita denominado espacio de reducción efectiva de la dimensión. La estimación de este modelo se lleva a cabo mediante la estimación de la función de regresión inversa  $\mathbb{E}(Y \nearrow X)$ . Para ello, este método requiere de algunas restricciones en la distribución de Y (*e.g.*, que Y tenga una distribución elíptica).

A pesar del creciente trabajo en el campo funcional, subsisten varios problemas abiertos. Uno de ellos es que ninguno de los métodos no paramétricos existentes conduce a la reducción de la dimensión. En este sentido resulta interesante explorar las posibilidades de extender el enfoque de Máquinas de vectores soportes, una bien conocida técnica de reducción de la dimensión para datos multivariados, a datos funcionales. Otro problema abierto es que ninguno de los métodos no paramétricos funcionales existentes hacen completo uso de la densidad condicional  $f(\cdot; X, \boldsymbol{\theta})$ , es decir, no incorporan una exhaustiva información del modelo, que en el caso de espectrometría, es el mecanismo físico generador de los datos. El método FIR sólo usa información parcial (momentos de segundo orden) del modelo y, además, impone restricciones fuertes en la distribución de Y.

El objetivo general del presente trabajo de tesis es desarrollar nuevos métodos de calibración para datos con valores en diversos espacios muestrales, univariados, multivariados y funcionales que mitiguen algunas de las deficiencias de los enfoques existentes. Para ello se proponen los siguientes objetivos específicos:

- Elaborar un enfoque predictivo no Bayesiano (frecuentista) para la calibración univariada lineal que brinde estimadores con buenas propiedades, tanto asintóticas como para muestras finitas, y que además de estimaciones puntuales, permita explorar las verosimilitudes de los posibles valores del parámetro de interés mediante una verosimilitud predictiva.
- 2. Evaluar la utilidad del enfoque de Máquinas de vectores relevantes para problemas de calibración multivariada, y explorar su comportamiento en datos reales de espectrometría.
- 3. Extender el método Máquinas de vectores soportes a datos funcionales para su aplicación a problemas de calibración con datos funcionales.
- 4. Elaborar un método de regresión inversa para datos funcionales que haga completo uso de la densidad condicional asociada al modelo de calibración, que no requiera de supuestos tan restrictivos en comparación con el método FIR y sea computacionalmente simple.

La tesis está estructurada en introducción, cuatro capítulos, conclusiones, recomendaciones y anexos.

En el Capítulo 1 se desarrolla un nuevo enfoque predictivo no Bayesiano (frecuentista) para la calibración univariada lineal. El mismo comienza con una revisión detallada de los principales métodos que se han desarrollado hasta el momento en esta temática. La nueva propuesta que se presenta en este capítulo supera las deficiencias del enfoque clásico (de verosimilitud). De forma similar a los métodos Bayesianos (y contrariamente al estimador clásico) tiene en consideración (a través de la construcción de una densidad predictiva) la incertidumbre estadística sobre los parámetros del modelo. Brinda no sólo un estimador puntual  $\hat{x}_P$  de  $x_0$ , sino que además permite explorar la verosimilitud (predictiva) para cualquier valor posible de  $x_0$ . A diferencia de los métodos Bayesianos (y al igual que el clásico), no requiere de especificaciones de distribuciones a priori para los parámetros del modelo, ni tampoco que los puntos de diseño  $x_1, x_2, ..., x_n$  y el nuevo punto  $x_0$  sean generados por una densidad de probabilidad común, lo cual lo hace un enfoque potencialmente útil para los casos de calibración controlada, donde el problema de calibración pueda involucrar extrapolación. Además, se estudian algunas propiedades del estimador como son la finitud de sus momentos, su consistencia y otras propiedades asintóticas. Finalmente, se muestran los resultados de un estudio de simulación donde se explora el comportamiento de la nueva propuesta, así como su comparación con algunos de los métodos reportados en la literatura. Los resultados expuestos en este capítulo se encuentran aceptados para publicar en [42].

En el Capítulo 2 se aplican por primera vez las Máquinas de vectores relevantes para el problema de la calibración multivariada. Este comienza con una breve descripción de los métodos disponibles para la calibración multivariada, sus ventajas y limitaciones. Luego se expone el marco teórico de las RVMR, y finalmente se hace un estudio de su comportamiento en tres conjuntos de datos reales de espectrometría. Los resultados expuestos se encuentran publicados en [47, 72].

El Capítulo 3 se dedica a la extensión del método Máquinas de vectores soportes a datos funcionales, que da lugar a un nuevo método denominado Máquinas de Vectores Soportes Funcionales (FSVR, del inglés <u>Functional Support Vector Regression</u>). Esta nueva propuesta, a diferencia de los métodos no paramétricos existentes, conduce a la reducción de la dimensión y no requiere supuestos tan restrictivos como los del método FIR, enfoque semi-paramétrico que conduce también a la reducción de la dimensión. El capítulo se inicia con una revisión detallada de los principales métodos desarrollados en la temática, y se precisan sus ventajas y desventajas. Se presenta la fundamentación teórica del método en el marco de la aproximación de funciones en RKHS, así como una nueva clase de funciones núcleos aplicable a datos funcionales. Finalmente se evalúa su comportamiento en un conjunto de datos reales de espectrometría. Los resultados presentados se publicaron en [41, 46].

En el Capítulo 4 se elabora un método de regresión inversa para datos funcionales aplicable a la calibración funcional. La nueva propuesta incorpora información sobre el modelo físico generador de los datos y no requiere supuestos tan restrictivos como los de FIR. Esta propuesta se denomina Regresión Inversa Basada en la Densidad (DBIR, del inglés <u>Density-Based Inverse Regression</u>). DBIR, al igual que otros enfoques no paramétricos, permite estimar funciones de regresión no lineales. Este método requiere estimar el modelo inverso (como sucede en FIR), lo cual se lleva a cabo usando un método de regresión no paramétrico. Pero contrariamente a los métodos de regresión no paramétricos usuales en el FDA, esta función está definida en los reales (y no sobre un espacio de Hilbert de dimensión infinita), por lo que la tarea de estimación es más fácil, al igual que la selección de los posibles hiperparámetros involucrados en este proceso. Además, no requiere estimar un número grande de parámetros como es el caso de las redes neuronales funcionales. Por tanto, la nueva propuesta es computacionalmente fácil de implementar. A diferencia del enfoque FIR, son necesarios pocos supuestos: en particular, la función de regresión no necesita ser función de un número finito de proyecciones, tampoco Y tiene que seguir una distribución elíptica (o cualquier otra distribución) y no requiere el último paso de estimación no paramétrica multivariada que se requiere en FIR. Su comportamiento se explora a través de un estudio de simulaciones y su aplicación se ilustra a través de un conjunto de datos reales de espectrometría. Parte de los resultados de este capítulo se encuentran aceptados para su publicación en [43] y otros están enviados a publicar [44].

# Capítulo 1

# UN ENFOQUE PREDICTIVO NO BAYESIANO PARA LA CALIBRACIÓN UNIVARIADA

### 1.1. Revisión de los enfoques previos

En este capítulo se trata el caso en que la variable Y involucrada en el problema de calibración descrito en la Introducción toma valores en los reales, es decir, la calibración univariada.

Sea  $\mathcal{D}$  una muestra de observaciones independientes  $(x_1, y_1), ..., (x_n, y_n)$  de (X, Y)(muestra de calibración) y  $y_0$  una nueva observación de Y que corresponde a un valor desconocido  $x_0$  de X. Si  $x_0$  está dentro del rango de los datos de calibración  $(x_1, ..., x_n)$ , se dice que el problema de calibración es un problema de interpolación; de lo contrario, se dice que es un problema de extrapolación. En algunos experimentos, se tienen varias observaciones  $y_{01}, ..., y_{0m}$  (m > 1) de Y que corresponden al mismo valor desconocido  $x_0$  de X. En este capítulo se hace énfasis en el caso en que sólo se obtiene una nueva observación  $y_0 = y_{01}$  (i.e., m = 1). Además, se restringe la atención a situaciones en que los valores  $x_i$  de la variable X en la muestra de calibración se fijan por el investigador (i.e., experimentos controlados). De manera más específica, se supone que el modelo estadístico  $f(\cdot; X, \boldsymbol{\theta})$  es un modelo de regresión lineal univariado Gaussiano:

$$Y = \alpha + \beta X + \epsilon, \tag{1.1}$$

donde la perturbación  $\epsilon$  tiene una distribución normal  $N(0, \sigma^2)$ , y el vector (fila) de

parámetros desconocidos es  $\boldsymbol{\theta} = (\alpha, \beta, \sigma^2).$ 

Para este problema de calibración se han propuesto varios enfoques, los cuales pueden ser agrupados en tres clases fundamentales:

#### a) Estimador máximo verosímil o estimador clásico

Al aplicar el método de máxima verosimilitud para estimar los parámetros desconocidos  $\boldsymbol{\theta}$  y  $x_0$  se obtiene lo que se conoce como estimador clásico  $\hat{x}_C$  de  $x_0$ , introducido por Eisenhart [24]:

$$\hat{x}_C = \frac{y_0 - \hat{\alpha}}{\hat{\beta}} = \overline{x} + \frac{1}{\hat{\beta}}(y_0 - \overline{y}), \qquad (1.2)$$

donde  $\hat{\alpha}, \hat{\beta}$  son las estimaciones máximo verosímiles (*i.e.*, de mínimos cuadrados) de los coeficientes de regresión  $\alpha, \beta$  a partir de los datos  $\mathcal{D}, y \, \overline{x}, \overline{y}$  son los valores de las medias de  $x_i y y_i$  (i = 1, ..., n), respectivamente. El estimador  $\hat{x}_C$  puede obtenerse también invirtiendo la recta de regresión ajustada  $y_0 = \hat{\alpha} + \hat{\beta}x_0$  con respecto a  $x_0$ . Si se tuvieran varias observaciones futuras  $y_{01}, ..., y_{0m}$  (m > 1) de Y, entonces  $\hat{x}_C$  se define reemplazando  $y_0$  en la expresión (1.2) por el valor de la media  $\overline{y}_0$  de  $y_{01}, ..., y_{0m}$ .

Es conocido que el estimador clásico es consistente; *i.e.*,  $\hat{x}_C$  converge a  $x_0$  cuando n y m tienden a infinito. También, su error cuadrático medio asintótico es la función constante  $\sigma^2/\beta^2$  en todo el rango de posibles valores de  $x_0$  [4]. Sin embargo, este estimador ha sido muy criticado porque no tiene momentos finitos, y por tanto su error cuadrático medio es infinito [105].

#### b) Estimadores Bayesianos

Como alternativas al estimador clásico, se ha propuesto un número de estimadores que pueden justificarse desde una perspectiva Bayesiana. En particular, este es el caso del estimador inverso  $\hat{x}_I$  propuesto por Krutchkoff [52, 53]. Krutchkoff construyó este estimador considerando el modelo de regresión lineal de X como función de Y (*i.e.*, la regresión "inversa" en comparación con hacer la regresión de Y versus X), es decir

$$X = \gamma + \delta Y + \eta,$$

calculó las estimaciones por mínimos cuadrados  $\hat{\gamma}$  y  $\hat{\delta}$  de los coeficientes  $\gamma$  y  $\delta$  de este modelo, y evaluó la recta de regresión en  $Y = y_0$  para obtener

$$\hat{x}_I = \hat{\gamma} + \hat{\delta} y_0.$$

Esto puede expresarse también como

$$\hat{x}_I = \overline{x} + \frac{\hat{\beta}}{\hat{\beta}^2 + \frac{n-2}{n} \frac{\hat{\sigma}^2}{\sigma_x^2}} (y_0 - \overline{y}), \qquad (1.3)$$

donde

$$\hat{\sigma}^2 = \frac{1}{n-2} \sum (y_i - \hat{y}_i)^2, \qquad (1.4)$$

con

$$\hat{y}_i = \hat{\alpha} + \beta x_i,$$
  
 $\sigma_x^2 = \frac{1}{n} \sum (x_i - \overline{x})^2$ 

El estimador inverso puede obtenerse también como un estimador Bayesiano bajo distribuciones a priori no informativas [1]. Alternativamente, puede derivarse minimizando (sobre cierta clase de estimadores lineales) versiones empíricas del riesgo Bayesiano, como son: el criterio de validación cruzada, definido por Perng [71] o el criterio del error compuesto, definido por Lwin y Maritz [60].

Varios estimadores se reducen o son muy similares al estimador inverso cuando el número m de observaciones de Y correspondiente al valor desconocido  $X = x_0$ es m = 1. Este es el caso de los estimadores propuestos por Halperin [36] y por Hagwood [35].

En efecto, el estimador de Halperin [36] está dado por

$$\hat{x}_H = (1 - R_M)\overline{x} + R_M \hat{x}_C, \qquad (1.5)$$

donde

$$R_M = \frac{\hat{\beta}^2}{\frac{n-2}{nM}\frac{\hat{\sigma}^2}{\sigma_x^2}}$$

y M es una función específica de m. La selección usual M = m conduce a  $\hat{x}_H = \hat{x}_I$  cuando m = 1, como puede apreciarse de manera directa comparando las expresiones (1.5) y (1.3).

Hagwood [35] consideró el problema de calibración como un problema inverso mal condicionado, cuya solución sobre la base de la regularización cuadrática conduce a la familia de estimadores

$$\hat{x}_{\lambda} = \overline{x} + \frac{\hat{\beta}}{\hat{\beta}^2 + \lambda} (y_0 - \overline{y}),$$

que depende del parámetro de regularización  $\lambda > 0$ . Demostró además que la esperanza de la función de validación cruzada asintótica se minimiza tomando  $\lambda = \hat{\lambda}_{cv} = \hat{\sigma}^2 / \sigma_x^2$ , lo cual da como resultado el estimador

$$\hat{x}_{\hat{\lambda}_{cv}} = \overline{x} + \frac{\hat{\beta}}{\hat{\beta}^2 + \frac{\hat{\sigma}^2}{\sigma_x^2}} (y_0 - \overline{y}).$$

Nótese que este estimador es prácticamente indistinguible de  $\hat{x}_I$  (comparar con la ecuación (1.3)), especialmente cuando n es moderada o grande.

c) Estimadores basados en correcciones asintóticas al estimador clásico

En la literatura se reporta un número de correcciones o modificaciones al estimador clásico.

Algunas de ellas consisten en sumas ponderadas de estimadores. Por ejemplo, sobre la base de análisis asintótico con respecto al tamaño de muestra, Ali y Singh [2] propusieron la siguiente suma ponderada del estimador clásico  $\hat{x}_C$  y  $\overline{x}$ :

$$\hat{x}_{AS} = \lambda \hat{x}_C + (1 - \lambda)\overline{x}$$

donde

$$\hat{\lambda} = \frac{\hat{\beta}^2 \hat{\zeta}^2}{\hat{\beta}^2 \hat{\zeta}^2 + \hat{\sigma}^2}$$

y  $\hat{\zeta} = \hat{x}_C - \overline{x}$  o  $\hat{\zeta} = \hat{x}_I - \overline{x}$ . La idea subyacente es mejorar el comportamiento del estimador  $\hat{x}_C$  en la interpolación, contrayéndolo hacia  $\overline{x}$  cuando  $x_0$  está cercano a  $\overline{x}$ . Tal contracción resulta ser menor cuando  $x_0$  está alejado de  $\overline{x}$ , y por tanto se espera que  $\hat{x}_{AS}$  retenga el buen comportamiento de  $\hat{x}_C$  para la extrapolación.

Otro estimador de suma ponderada fue propuesto por Srivastava y Singh [87] sobre la base de la teoría asintótica con respecto a perturbaciones pequeñas:

$$\hat{x}_{SS} = (1 - \lambda)\hat{x}_C + \lambda\hat{x}_I, \tag{1.6}$$

donde  $\lambda = 1/(n-2)$ . Este estimador puede considerarse como una corrección leve al estimador inverso  $\hat{x}_I$  cuando el tamaño de muestra *n* es pequeño, y una corrección leve al estimador clásico  $\hat{x}_C$  cuando n es grande.

En aras de garantizar insesgadez asintótica, Naszódi [68] propuso otra corrección al estimador clásico que conduce al estimador:

$$\hat{x}_N = \overline{x} + \frac{\hat{\beta}}{\hat{\beta}^2 + \frac{1}{n}\frac{\hat{\sigma}^2}{\sigma_x^2}}(y_0 - \overline{y}).$$
(1.7)

Un número de estudios teóricos y de simulaciones publicados comparan los estimadores clásico e inverso con respecto a varios criterios ([4, 16, 17, 51, 52, 53, 60, 84, 99, 93, 105]). Los resultados principales de esta comparación pueden resumirse como sigue:

- (I) El estimador clásico muestra mejor comportamiento que el estimador inverso para la extrapolación y peor comportamiento para la interpolación, tanto para tamaños de muestra moderados como grandes. En particular,  $\hat{x}_I$  tiene error cuadrático medio asintótico no acotado como función de  $x_0$  (específicamente, es una función creciente de  $(x_0 - \overline{x})^2$ ) mientras que  $\hat{x}_C$  tiene error cuadrático medio asintótico constante. Además, el mínimo de la función de error cuadrático medio asintótica de  $\hat{x}_I$ , el cual se alcanza en  $x_0 = \overline{x}$ , es menor que el de  $\hat{x}_C$ .
- (II)  $\hat{x}_C$  es un estimador consistente, mientras  $\hat{x}_I$  no lo es.
- (III) Al contrario de  $\hat{x}_C$ ,  $\hat{x}_I$  tiene momentos finitos, y por tanto MSE finito.
- (IV) El comportamiento de  $\hat{x}_C$  con muestras finitas se deteriora de forma más drástica que el de  $\hat{x}_I$ , a medida que  $\beta$  se vuelve pequeño y  $\sigma^2$  grande.

Puede concluirse que los enfoques no Bayesianos existentes para el problema de la calibración univariada poseen propiedades estadísticas marcadamente indeseables, como es el caso del estimador clásico, o sus virtudes están solamente fundamentadas por consideraciones asintóticas, no válidas para muestras finitas. Además, sólo brindan estimaciones de tipo puntual o por intervalos.

A diferencia de los métodos descritos anteriormente, el enfoque que se presenta en este capítulo para la calibración no se basa en máxima verosimilitud, teoría Bayesiana o correcciones asintóticas sino en un marco predictivo no Bayesiano. Un enfoque predictivo es muy apropiado para la calibración porque esta constituye precisamente un problema de estadística predictiva: dado una muestra previa de entrenamiento, interesa hacer cierto tipo de inferencia para una muestra futura. De esta forma se le da solución al problema de elaborar un enfoque no Bayesiano (frecuentista) para el problema de la calibración univariada lineal que mitigue las deficiencias anteriormente mencionadas, como se verá en este capítulo.

Para esto, en la sección siguiente se presenta la densidad predictiva no Bayesiana introducida por Harris [37] generalizada al contexto de calibración, de esta forma se brinda una densidad predictiva  $f_P(\cdot; x_0, \mathcal{D})$  para una observación y correspondiente a  $X = x_0$ . Sobre la base de este resultado, se define un estimador no Bayesiano  $\hat{x}_P$ de  $x_0$ . Este enfoque alternativo tiene un número de características promisorias que motivan su investigación:

- (I) Al igual que el estimador clásico y a diferencia de los estimadores Bayesianos, no requiere especificación de alguna distribución a priori para los parámetros del modelo, ni el supuesto de que los puntos de diseño  $x_1,...,x_n$  y el nuevo punto  $x_0$  sean generados por una densidad de probabilidad común. Esto último hace este nuevo enfoque potencialmente útil para situaciones en las cuales los datos  $x_i$  provienen de experimentos controlados y el problema de calibración será del tipo extrapolación.
- (II) Similar a los métodos Bayesianos y contrario al estimador clásico, tiene en consideración (a través de la construcción de la densidad predictiva  $f_P$ ) la incertidumbre estadística sobre los parámetros del modelo  $\alpha, \beta, \sigma^2$ . Es de esperar que esto tenga un efecto de regularización, y por tanto la precisión de  $\hat{x}_P$  no se deteriore tanto como la de  $\hat{x}_C$  cuando el tamaño de muestra no es grande y  $\beta$  es pequeño y  $\sigma^2$  grande.
- (III) Permite no sólo obtener estimaciones puntuales  $\hat{x}_P$  de  $x_0$  sino explorar la verosimilitud (predictiva) de cualquier valor posible de  $x_0$ .

### 1.2. Estimador predictivo no Bayesiano

Sea  $\mathcal{Y}$  una muestra de observaciones independientes e idénticamente distribuidas  $y_1, ..., y_n$  cuya densidad  $f(\cdot; \boldsymbol{\theta}_0)$  pertenece a una familia de densidades  $f(\cdot; \boldsymbol{\theta})$ , que dependen de un vector de parámetros desconocidos  $\boldsymbol{\theta} \in \Theta$ . Sobre la base de los datos  $\mathcal{Y}$ , Harris [37] definió la siguiente densidad predictiva no Bayesiana para una observación futura Y de la misma distribución  $f(\cdot; \boldsymbol{\theta}_0)$ :

$$f_P(y; \mathcal{Y}) = \int f(y; \boldsymbol{\theta}) \hat{f}_{\hat{\boldsymbol{\theta}}}(\boldsymbol{\theta}) d\boldsymbol{\theta},$$

donde  $\hat{\boldsymbol{\theta}}$  es la estimación máximo verosímil de  $\boldsymbol{\theta}_0$  sobre la base de  $(y_1, ..., y_n)$ , y  $\hat{f}_{\hat{\boldsymbol{\theta}}}$  es una estimación de la densidad  $f_{\hat{\boldsymbol{\theta}}}(\cdot; \boldsymbol{\theta}_0)$  para  $\hat{\boldsymbol{\theta}}$ . Usualmente,  $\hat{f}_{\hat{\boldsymbol{\theta}}}$  es una estimación por sustitución (<u>plug-in</u>) resultante de reemplazar  $\boldsymbol{\theta}_0$  por  $\hat{\boldsymbol{\theta}}$  en las expresiones analíticas de la densidad de  $\hat{\boldsymbol{\theta}}$ ; *i.e.*,  $\hat{f}_{\hat{\boldsymbol{\theta}}}(\boldsymbol{\theta}) = f_{\hat{\boldsymbol{\theta}}}(\cdot; \hat{\boldsymbol{\theta}})$ . Nótese que la idea esencial de tal densidad predictiva es ponderar los valores posibles de  $\boldsymbol{\theta}$  de acuerdo con la información brindada por la densidad de su estimador máximo verosímil sobre la base de la muestra de observaciones pasadas.

A continuación, se propone una generalización de esta definición a modelos de regresión. Específicamente, para i = 1, ..., n, sea  $y_i$  una observación de una densidad  $f(\cdot; x_i, \boldsymbol{\theta}_0)$  que depende del parámetro desconocido  $\boldsymbol{\theta}_0$  y también del valor  $x_i$ de alguna variable X, y se denota por  $\mathcal{D}$ , los datos (pasados) de entrenamiento  $(x_1, y_1), ..., (x_n, y_n)$ . Una densidad predictiva no Bayesiana para una observación futura Y de la densidad  $f(\cdot; x_0, \boldsymbol{\theta}_0)$  correspondiente al valor de la covariable  $X = x_0$ puede definirse como

$$f_P(y; x_0, \mathcal{D}) = \int f(y; x_0, \boldsymbol{\theta}) \hat{f}_{\hat{\boldsymbol{\theta}}}(\boldsymbol{\theta}) d\boldsymbol{\theta}, \qquad (1.8)$$

donde  $\hat{\boldsymbol{\theta}}$  es la estimación máximo verosímil de  $\boldsymbol{\theta}_0$  basada en los datos  $\mathcal{D}$ , y  $\hat{f}_{\hat{\boldsymbol{\theta}}}$  es una estimación de la densidad  $f_{\hat{\boldsymbol{\theta}}}(\cdot;\boldsymbol{\theta}_0)$  de  $\hat{\boldsymbol{\theta}}$ .

Este concepto será aplicado aquí para construir una densidad predictiva no Bayesiana para el problema de calibración. Para esto, es conveniente reescribir el modelo (1.1) en forma vectorial

$$Y = \mathbf{u}\boldsymbol{\gamma}^{\top} + \boldsymbol{\epsilon},$$

donde  $\mathbf{u} = (1, X)$  y  $\boldsymbol{\gamma} = (\alpha, \beta)$  son vectores filas, y  $\top$  denota el operador de transposición de matrices. El vector de parámetros de este modelo es  $\boldsymbol{\theta} = (\boldsymbol{\gamma}, \sigma^2)$ . El valor verdadero del vector de parámetros se denota por  $\boldsymbol{\theta}_0 = (\boldsymbol{\gamma}_0, \sigma_0^2)$ , donde  $\boldsymbol{\gamma}_0 = (\alpha_0, \beta_0)$ . Se usan las estimaciones  $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\gamma}}, \hat{\sigma}^2)$ , donde  $\hat{\boldsymbol{\gamma}} = (\hat{\alpha}, \hat{\beta})$  es la estimación máximo verosímil (*i.e.*, por mínimos cuadrados) de los coeficientes del modelo  $\alpha$  y  $\beta$ , y  $\hat{\sigma}^2$  es la estimación insesgada dada por (1.4).

Se conoce que la densidad de  $\hat{\theta}$  es

$$\begin{split} f_{\hat{\boldsymbol{\theta}}}(\boldsymbol{\theta};\boldsymbol{\theta}_0) &= f_{(\hat{\boldsymbol{\gamma}},\hat{\sigma}^2)}(\boldsymbol{\gamma},\sigma^2;\boldsymbol{\gamma}_0,\sigma_0^2) \\ &= f_{\hat{\boldsymbol{\gamma}}}(\boldsymbol{\gamma};\boldsymbol{\gamma}_0,\sigma_0^2) f_{\hat{\sigma}^2}(\sigma^2;\sigma_0^2), \end{split}$$

donde las densidades de  $\hat{\boldsymbol{\gamma}}$  y  $\hat{\sigma}^2$  son, respectivamente:

$$\begin{split} f_{\hat{\boldsymbol{\gamma}}}(\boldsymbol{\gamma};\boldsymbol{\gamma}_{0},\sigma_{0}^{2}) &= f_{N\left(\boldsymbol{\gamma}_{0},\left(\mathbf{U}^{\top}\mathbf{U}\right)^{-1}\sigma_{0}^{2}\right)}(\boldsymbol{\gamma}),\\ f_{\hat{\sigma}^{2}}(v;\sigma_{0}^{2}) &= \frac{l}{\sigma_{0}^{2}}f_{\chi^{2}_{(l)}}\left(\frac{l}{\sigma_{0}^{2}}v\right). \end{split}$$

Aquí,  $f_{N(\boldsymbol{\gamma}_{0}, (\mathbf{U}^{\top}\mathbf{U})^{-1}\sigma_{0}^{2})}$  es la densidad Gaussiana (bivariada) con media  $\boldsymbol{\gamma}_{0}$  y matriz de covarianza  $(\mathbf{U}^{\top}\mathbf{U})^{-1}\sigma_{0}^{2}$ ;  $\mathbf{U}$  es la matriz de dimensión  $n \times 2$  con filas  $\mathbf{u}_{i} = (1, x_{i})$ , i = 1, ..., n; y  $f_{\chi^{2}_{(l)}}$  es la densidad chi-cuadrado con l = n - 2 grados de libertad. Por consiguiente, una estimación por sustitución de  $f_{\hat{\theta}}(\boldsymbol{\theta}; \boldsymbol{\theta}_{0})$  está dada por

$$\hat{f}_{\hat{\boldsymbol{\theta}}}(\boldsymbol{\theta}) = \hat{f}_{(\hat{\boldsymbol{\gamma}},\hat{\sigma}^2)}(\boldsymbol{\gamma},\sigma^2;\hat{\boldsymbol{\gamma}},\hat{\sigma}^2) = f_{\hat{\boldsymbol{\gamma}}}(\boldsymbol{\gamma};\hat{\boldsymbol{\gamma}},\hat{\sigma}^2)f_{\hat{\sigma}^2}(\sigma^2;\hat{\sigma}^2).$$

Tomando en consideración que  $f(y; x_0, \boldsymbol{\theta}) = f_{N(\mathbf{u}_0 \boldsymbol{\gamma}^{\top}, \sigma^2)}(y)$ , donde  $\mathbf{u}_0 = (1, x_0)$ , y

$$\int f_{N\left(\mathbf{u}_{0}\boldsymbol{\gamma}^{\top},\sigma^{2}\right)}(y)f_{\hat{\boldsymbol{\gamma}}}(\boldsymbol{\gamma};\hat{\boldsymbol{\gamma}},\hat{\sigma}^{2})d\boldsymbol{\gamma} = f_{N\left(\mathbf{u}_{0}\hat{\boldsymbol{\gamma}}^{\top},\mathbf{u}_{0}\left(\mathbf{U}^{\top}\mathbf{U}\right)^{-1}\mathbf{u}_{0}^{\top}\hat{\sigma}^{2}+\sigma^{2}\right)}(y), \qquad (1.9)$$

la densidad predictiva (1.8) se reduce a

$$f_P(y; x_0, \mathcal{D}) = \int_0^\infty f_{N\left(\mathbf{u}_0 \hat{\boldsymbol{\gamma}}^\top, \mathbf{u}_0 \left(\mathbf{U}^\top \mathbf{U}\right)^{-1} \mathbf{u}_0^\top \hat{\sigma}^2 + v\right)}(y) f_{\hat{\sigma}^2}(v; \hat{\sigma}^2) dv.$$
(1.10)

La verosimilitud de diferentes valores posibles de  $x_0$  puede obtenerse a través de lo que se denomina función de verosimilitud predictiva para  $x_0$ . Esta se define como la función  $L_P(x_0) = f_P(y_0; x_0, \mathcal{D})$  sobre el rango de posibles valores de  $x_0$ , para los datos  $y_0$  y  $\mathcal{D}$ .

Además, se define el estimador predictivo no Bayesiano  $\hat{x}_P$  de  $x_0$  maximizando la función de verosimilitud predictiva, esto es,

$$\hat{x}_P = \arg \max_{x \in \mathbb{R}} L_P(x). \tag{1.11}$$

Nótese que estas definiciones pueden extenderse al caso en que se tienen varias observaciones  $y_{01}, ..., y_{0m}$  (m > 1) de  $Y_0$ . En tal caso la densidad predictiva para la media  $\overline{y}_0$  de las observaciones (futuras)  $y_{01}, ..., y_{0m}$  está dada por

$$L_P(x_0) = f_P(\overline{y}_0; x_0, \mathcal{D}) = \int_0^\infty f_{N(\mathbf{u}_0 \hat{\boldsymbol{\gamma}}^{\top}, \mathbf{u}_0 (\mathbf{U}^{\top} \mathbf{U})^{-1} \mathbf{u}_0^{\top} \hat{\sigma}^2 + v/m)}(\overline{y}_0) f_{\hat{\sigma}^2}(v; \hat{\sigma}^2) dv, \quad (1.12)$$

y el estimador predictivo se obtiene maximizando esta función con respecto a  $x_0$ .

# 1.3. Propiedades del estimador predictivo no Bayesiano

En esta sección se estudian algunas propiedades del estimador predictivo no Bayesiano. En primer lugar se obtiene una forma explícita (aproximada) para el estimador a través del desarrollo asintótico del mismo. Partiendo de esta expresión, se obtienen los momentos de primero y segundo órdenes asintóticos, con lo cual se calcula su error cuadrático medio asintótico. Luego se demuestra la finitud de sus momentos y por último, se estudia su consistencia.

En las secciones que siguen, el símbolo " $\approx$ " expresa que el miembro derecho de las ecuaciones se aproxima a su miembro izquierdo, excepto para términos residuales  $r_n$  de orden  $1/n^2$ ; y esto significa que  $r_n = O_P(1/n^2)$  o  $r_n = O(1/n^2)$  en dependencia de si el desarrollo de Taylor es en probabilidad o no. Detalles técnicos adicionales sobre desarrollos de Taylor en probabilidad pueden encontrarse en [83].

#### 1.3.1. Desarrollo asintótico del estimador

**Teorema 1.1.** El estimador predictivo no Bayesiano  $\hat{x}_P$ , definido en (1.11), reteniendo términos de hasta aproximación asintótica de primer orden con respecto a  $\varepsilon = 1/n$ puede ser aproximado por

$$\hat{x}_P \approx \overline{x} + \frac{(1 - \hat{\eta}_n^2)}{\hat{\beta}} \left( y_0 - \overline{y} \right), \qquad (1.13)$$

donde

$$\hat{\eta}_n^2 = \frac{\hat{\sigma}^2}{\sigma_x^2 \hat{\beta}^2 n}.$$
(1.14)

Demostración:

Sea  $\hat{\mu}_0 = \mathbf{u}_0 \hat{\boldsymbol{\gamma}}^\top = \hat{\alpha} + \hat{\beta} x_0, w_0 = \mathbf{u}_0 \left( \mathbf{U}^\top \mathbf{U} \right)^{-1} \mathbf{u}_0^\top \text{ y } G = -\ln f_P(y_0; x_0, \mathcal{D}).$  Al hacer el desarrollo de Taylor de  $f_{N(\hat{\mu}_0, w_0 \hat{\sigma}^2 + v/m)}(y_0)$  (definida por la expresión (1.10)) con respecto a v alrededor de  $\hat{\sigma}^2$  y tomar en consideración que  $(n-2)\hat{\sigma}^2 \neq \sigma^2$  es una variable aleatoria con distribución chi-cuadrado con n-2 grados de libertad se tiene que

$$f_P(y_0; x_0, \mathcal{D}) = f_{N(\hat{\mu}_0, w_0 \hat{\sigma}^2 + \hat{\sigma}^2)}(y_0) + \frac{\partial^2}{\partial v^2} f_{N(\hat{\mu}_0, w_0 \hat{\sigma}^2 + v)}(y_0) \big|_{v = \hat{\sigma}^2} \frac{\hat{\sigma}^4}{n - 2} + R,$$

donde el término remanente R es de orden  $1/n^2$ . Esto y

$$w_0 = \frac{(x_0 - \overline{x})^2}{\sigma_x^2 (n - 1)} + \frac{1}{n}$$

implican que, hasta el desarrollo de primer orden con respecto a $\varepsilon=1/n,$  se cumple que

$$G \approx \frac{1}{2} \ln (2\pi) + \frac{(y_0 - \hat{\mu}_0)^2}{2\hat{\sigma}^2 \left(\varepsilon \frac{(x_0 - \overline{x})^2}{\sigma_x^2} + \varepsilon + 1\right)} + \frac{1}{2} \ln \left( \left( \frac{(x_0 - \overline{x})^2}{\sigma_x^2} \varepsilon + \varepsilon + 1 \right) \hat{\sigma}^2 \right)$$
(1.15)  
$$- \frac{3 (y_0 - \hat{\mu}_0)^2}{2 \left(\varepsilon \frac{(x_0 - \overline{x})^2}{\sigma_x^2} + \varepsilon + 1\right)^3 \hat{\sigma}^2} \varepsilon + \frac{(y_0 - \hat{\mu}_0)^4}{4\hat{\sigma}^4 \left(\varepsilon \frac{(x_0 - \overline{x})^2}{\sigma_x^2} + \varepsilon + 1\right)^4} \varepsilon + \frac{3}{4 \left(\varepsilon \frac{(x_0 - \overline{x})^2}{\sigma_x^2} + \varepsilon + 1\right)^2} \varepsilon.$$

Por definición,  $\hat{x}_P$  minimiza a G. De este modo,

$$\frac{\partial}{\partial x_0} G \Big|_{x_0 = \hat{x}_P} = 0 \tag{1.16}$$

para cualquier  $\varepsilon > 0$ . En  $\varepsilon = 0$  se tiene

$$G|_{\varepsilon=0} = -\ln f_{N(\hat{\mu}_0,\hat{\sigma}^2)}(y_0),$$

la cual es una función minimizada por  $\hat{x}_C$ ; por tanto  $\hat{x}_P|_{\varepsilon=0} = \hat{x}_C$ . Por consiguiente, el desarrollo de Taylor de primer orden de  $\hat{x}_P$  con respecto a  $\varepsilon$  queda

$$\hat{x}_P \approx \hat{x}_C + \frac{\partial}{\partial \epsilon} \hat{x}_P |_{\varepsilon=0} \varepsilon.$$
 (1.17)

Por otra parte, derivando la ecuación (1.16) con respecto a  $\varepsilon$  se tiene que

$$\frac{\partial}{\partial \epsilon} \hat{x}_P |_{\epsilon=0} = -\frac{\frac{\partial}{\partial \epsilon} \frac{\partial}{\partial x_0} G \Big|_{x_0=\hat{x}_P}}{\frac{\partial^2}{\partial x_0^2} G \Big|_{x_0=\hat{x}_P}}$$
(1.18)

De (1.15), (1.17) y (1.18) sigue que, hasta la aproximación de primer orden con respecto a  $\varepsilon$ , el estimador predictivo no Bayesiano se aproxima por

$$\hat{x}_P \approx \overline{x} + \frac{(1 - \hat{\eta}_n^2)}{\hat{\beta}} (y_0 - \overline{y}),$$

donde  $\hat{\eta}_n^2$  está definido por (1.14), lo cual completa la demostración.  $\square$ 

### 1.3.2. Desarrollo asintótico del error cuadrático medio del estimador

**Teorema 1.2.** Reteniendo términos hasta orden  $\varepsilon = 1/n$ , el estimador predictivo no Bayesiano  $\hat{x}_P$  definido en (1.11), tiene error cuadrático medio asintótico

$$MSE(\hat{x}_{P}) = \mathbb{E}(\hat{x}_{P} - x_{0})^{2} \approx \frac{\sigma^{2}}{\beta^{2}} + \eta_{n}^{2} \left(\sigma_{x}^{2} + (x_{0} - \overline{x})^{2} + 3\frac{\sigma^{2}}{\beta^{2}}\right), \qquad (1.19)$$

donde

$$\eta_n^2 = \frac{\sigma^2}{\sigma_x^2 \beta^2 n}.\tag{1.20}$$

#### Demostración:

La aproximación del estimador predictivo no Bayesiano dada en (1.13) puede reexpresarse como

$$\hat{x}_P \approx \left(1 - \hat{\eta}_n^2\right) \frac{(y_0 - \hat{\alpha})}{\hat{\beta}} + \hat{\eta}_n^2 \overline{x}$$
(1.21)

donde  $\hat{\eta}_n^2$  está definido por (1.14).

Es necesario recordar que bajo el supuesto de errores Gaussianos en el modelo (1.1),  $(n-2)\hat{\sigma}^2 \swarrow \sigma^2$  es una variable aleatoria con distribución chi-cuadrado con n-2 grados de libertad, independiente del vector aleatorio  $(\hat{\alpha}, \hat{\beta})$ , y este último tiene distribución Gaussiana con media  $(\alpha, \beta)$  y matriz de covarianza  $\sigma^2 (\mathbf{U}^{\top} \mathbf{U})^{-1}$ .

Al calcular todas las derivadas hasta segundo orden de (1.21) con respecto a  $\hat{\alpha}, \hat{\beta}, \hat{\sigma}^2$  y evaluar el desarrollo de Taylor de la expresión (1.21) con respecto a  $(\hat{\alpha}, \hat{\beta}, \hat{\sigma})$  alrededor de  $(\alpha, \beta, \sigma)$  se obtiene

$$\hat{x}_P \approx A + C(\hat{\alpha} - \alpha) + D(\hat{\beta} - \beta) + E(\hat{\sigma}^2 - \sigma^2) + F(\hat{\beta} - \beta)^2$$

$$+ G(\hat{\alpha} - \alpha)(\hat{\beta} - \beta) + I(\hat{\beta} - \beta)(\hat{\sigma}^2 - \sigma^2) + H(\hat{\alpha} - \alpha)(\hat{\sigma}^2 - \sigma^2),$$
(1.22)

donde

$$A = (1 - \eta_n^2) \frac{(y_0 - \alpha)}{\beta} + \eta_n^2 \overline{x}$$

$$C = -\frac{(1 - \eta_n^2)}{\beta}$$

$$D = -(1 - \eta_n^2) \frac{(y_0 - \alpha)}{\beta^2} + 2 \frac{(y_0 - \alpha)}{\beta^4} \frac{\sigma^2}{\sigma_x^2 n} - 2\overline{x} \frac{\sigma^2}{\sigma_x^2 n \beta^3}$$

$$E = -\frac{(y_0 - \alpha)}{\beta^3 \sigma_x^2 n} + \frac{\overline{x}}{\beta^2 \sigma_x^2 n}$$

$$F = \frac{1}{2} \left[ -10 \frac{(y_0 - \alpha)\sigma^2}{\beta^5 \sigma_x^2 n} + 2 \frac{(y_0 - \alpha)(1 - \eta_n^2)}{\beta^3} + 6 \frac{\overline{x}\sigma^2}{\beta^4 \sigma_x^2 n} \right]$$
  

$$G = \frac{(1 - 3\eta_n^2)}{\beta^2}$$
  

$$I = 3 \frac{(y_0 - \alpha)}{\beta^4 \sigma_x^2 n} - 2 \frac{\overline{x}}{\beta^3 \sigma_x^2 n}$$
  

$$H = \frac{1}{\beta^3 \sigma_x^2 n}.$$

Apoyado en las fórmulas de Bartlett [54] para los momentos de variables aleatorias con distribución normal multivariada descritas en el siguiente Lema,

Lema 1.1. Si 
$$X = (X_1, ..., X_k) \sim N(\boldsymbol{\mu}, \Sigma) \ con \ \boldsymbol{\mu} = (\mu_1, ..., \mu_k) \ y$$
  

$$\Sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \dots & \sigma_{1k} \\ \vdots & \vdots & \ddots & \\ \sigma_{k1} & \sigma_{k2} & \dots & \sigma_{kk} \end{pmatrix} entonces$$

$$\mathbb{E}(x_i - \mu_i) = 0$$

$$\mathbb{E}(x_i - \mu_i)(x_j - \mu_j) = \sigma_{ij}$$

$$\mathbb{E}(x_i - \mu_i)(x_j - \mu_j)(x_p - \mu_p) = 0$$

$$\mathbb{E}(x_i - \mu_i)(x_j - \mu_j)(x_p - \mu_p) = \sigma_{ij} + \sigma_{ij}\sigma_{jl} + \sigma_{ij}\sigma_{jl} + \sigma_{il}\sigma_{jp},$$

y reteniendo sólo los términos hasta orden (1/n) en (1.22), se obtiene que

$$\mathbb{E}(\hat{x}_P) \approx x_0,\tag{1.23}$$

$$V(\hat{x}_P) \approx \frac{\sigma^2}{\beta^2} + \eta_n^2 \left( (x_0 - \overline{x})^2 - \frac{\sigma^2}{\beta^2} + \sigma_x^2 \right), \qquad (1.24)$$

$$MSE\left(\hat{x}_{P}\right) = \mathbb{E}\left(\hat{x}_{P} - x_{0}\right)^{2} \approx \frac{\sigma^{2}}{\beta^{2}} + \eta_{n}^{2}\left(\sigma_{x}^{2} + (x_{0} - \overline{x})^{2} + 3\frac{\sigma^{2}}{\beta^{2}}\right), \qquad (1.25)$$

donde  $\eta_n^2$  está definido por (1.20), con lo cual que<br/>da demostrado el teorema. $\square$ 

De estas aproximaciones se deduce un número de propiedades asintóticas del estimador predictivo  $\hat{x}_P$  que vale la pena mencionar:

1. Al retener sólo los términos hasta el orden cero en los desarrollos de Taylor con respecto a 1/n, el estimador predictivo  $\hat{x}_P$  coincide con el estimador clásico  $\hat{x}_C$  $(i.e., quitando el factor <math>\hat{\eta}_n^2 = O(1/n)$  en (1.13) se obtiene (1.2)). Por tanto, tienen las mismas propiedades asintóticas hasta el desarrollo de orden cero. Debido a esto, hasta el desarrollo de orden cero se cumple que: el MSE de  $\hat{x}_P$
es  $\sigma^2/\beta^2$  para todos los valores de  $x_0$ ; y bajo el criterio de MSE,  $\hat{x}_P$  (como el estimador clásico) es peor que el estimador inverso  $\hat{x}_I$  para la interpolación (*i.e.*, cuando  $x_0$  está lo suficientemente cercano a  $\overline{x}$ ), pero  $\hat{x}_P$  se comporta mejor que  $\hat{x}_I$  para la extrapolación (*i.e.*, cuando  $x_0$  se desvía lo suficiente de  $\overline{x}$ ).

2. Es conocido (ver, *e.g.*, [84]) que, hasta el desarrollo de primer orden, el *MSE* del estimador clásico está dado por

$$MSE(\hat{x}_C) \approx \frac{\sigma^2}{\beta^2} + \eta_n^2 \left[ (x_0 - \overline{x})^2 + 3\frac{\sigma^2}{\beta^2} + \sigma_x^2 \right].$$
(1.26)

Al comparar las ecuaciones (1.25) y (1.26) se tiene que, hasta el desarrollo de primer orden,

$$MSE(\hat{x}_C) - MSE(\hat{x}_P) \approx 4 \frac{\sigma^4}{\beta^4 \sigma_x^2 n} > 0$$

para cualquier valor de  $x_0$ . Por tanto, desde el punto de vista del MSE hasta el desarrollo de primer orden, el estimador propuesto  $\hat{x}_P$  es mejor que el estimador clásico  $\hat{x}_C$  en todo el rango de posibles valores de  $x_0$ . Además, nótese que  $MSE(\hat{x}_C) - MSE(\hat{x}_P)$  aumenta cuando  $\sigma^4$  se hace grande y  $\beta^4$ ,  $\sigma_x^2$  son pequeñas; es decir, la mejoría del nuevo estimador  $\hat{x}_P$  con respecto al clásico  $\hat{x}_C$ es mayor en situaciones en que la relación señal-ruido  $|\beta|/\sigma^2$  es pequeña.

3. Es conocido (ver, *e.g.*, [84]) que, hasta el desarrollo de Taylor de primer orden, el MSE de  $\hat{x}_I$  está dado por

$$MSE(\hat{x}_{I}) \approx \frac{\sigma^{2}}{\beta^{2} (1+\Psi)^{2}} \left[ \frac{1}{n} + 1 + \frac{(x_{0} - \overline{x})^{2}}{(n-1)\sigma_{x}^{2}} + \frac{\sigma^{2} (\Psi^{2} + 5)}{\beta^{2} (1+\Psi)^{2} (n-1)\sigma_{x}^{2}} + \frac{\sigma^{2} (x_{0} - \overline{x})^{2}}{\beta^{2} \sigma_{x}^{4}} \left( 1 - \frac{6}{n (1+\Psi)^{2}} \right) \right],$$

$$(1.27)$$

donde

$$\Psi = \frac{\sigma^2}{\beta^2 \sigma_x^2}$$

Cuando se comparan las expresiones (1.25) y (1.27) puede mostrarse que, hasta el desarrollo de primer orden,  $MSE(\hat{x}_I) < MSE(\hat{x}_P)$  cuando  $x_0 = \overline{x}$ . Esto significa que el estimador inverso se comporta mejor que  $\hat{x}_P$  para la interpolación. Puede mostrarse también que la diferencia  $MSE(\hat{x}_P) - MSE(\hat{x}_I)$  evaluada en  $x_0 = \overline{x}$  es una función creciente de  $\Psi$  (cuando n > 2). Esto significa que para  $x_0$  en una vecindad de  $\overline{x}$  el comportamiento de  $\hat{x}_P$  comparado con  $\hat{x}_I$  se deteriora de manera más notable cuando  $\sigma^2$  es grande y  $\beta^2$ ,  $\sigma_x^2$  son pequeñas.

Por otra parte, puede verse que para  $(x_0 - \overline{x})^2$  suficientemente grande se cumple que  $MSE(\hat{x}_I) > MSE(\hat{x}_P)$ . Es decir, para problemas de extrapolación,  $\hat{x}_P$  se comporta mejor que  $\hat{x}_I$ .

4. En el caso en que se tienen varias observaciones  $y_{01}, ..., y_{0m}$  (m > 1) de  $Y_0$  se cumple por (1.12) que, hasta el desarrollo de orden cero,  $\hat{x}_P \approx \overline{x} + (\overline{y}_0 - \overline{y})/\hat{\beta}$ . Esto implica que  $\hat{x}_P$  es un estimador de  $x_0$  consistente, *i.e.*,  $\hat{x}_P$  tiende a  $x_0$ cuando ambos,  $n \ge m$ , tienden a infinito.

A modo de conclusión se puede plantear que, hasta los desarrollos de Taylor de primer orden, el estimador  $\hat{x}_P$  propuesto se comporta mejor para la extrapolación que el estimador inverso y el estimador clásico. Para la interpolación, su comportamiento conduce a mejoras con respecto al estimador clásico aunque no mejora el del estimador inverso. Es importante señalar que la superioridad del estimador propuesto sobre el estimador clásico se hace mayor en situaciones en que  $\sigma^2$  es grande y  $\beta^2$  pequeña, es decir, cuando resulta más ventajoso regularizar a la hora de invertir el modelo de regresión para obtener  $x_0$ .

El comportamiento no asintótico del estimador predictivo no Bayesiano  $\hat{x}_P$  se estudia a través de simulaciones en secciones posteriores.

#### 1.3.3. Finitud de los momentos del estimador

Nótese que maximizar la densidad predictiva  $f_P(y_0; x, \mathcal{D})$  con respecto a x, involucra integrales con respecto a las densidades de  $\hat{\beta}$  y  $\hat{\sigma}^2$  (ver ecuaciones 1.9 y 1.10). En la práctica, realizar la integración con respecto a la densidad de  $\hat{\sigma}^2$  no tiene efectos numéricos apreciables en comparación con evaluar  $\sigma = \hat{\sigma}$ , *i.e.*, considerando  $f_{\hat{\sigma}^2}(\sigma^2; \hat{\sigma}^2)$ como una función delta de Dirac centrada en  $\sigma = \hat{\sigma}$ . De esta forma la densidad predictiva queda

$$\tilde{L}_P(x_0) = \tilde{f}_P(y_0; x_0, \mathcal{D}) = f_{N\left(\mathbf{u}_0 \hat{\boldsymbol{\gamma}}^\top, \mathbf{u}_0\left(\mathbf{U}^\top \mathbf{U}\right)^{-1} \mathbf{u}_0^\top \hat{\sigma}^2 + \hat{\sigma}^2\right)}(y)$$
(1.28)

y el estimador predictivo no Bayesiano resulta ser

$$\tilde{x}_P = \arg \max_{x \in \mathbb{R}} \tilde{L}_P(x).$$
(1.29)

Esta simplificación del estimador se tendrá en cuenta para los resultados que se presentan de aquí en adelante; es decir, para la demostración de momentos finitos, la demostración de consistencia y el estudio de simulación.

**Teorema 1.3.** Sea  $\tilde{x}_P$  el estimador predictivo no Bayesiano definido como (1.29). Entonces se cumple que

$$\mathbb{E}|\tilde{x}_P|^2 < \infty.$$

Demostración:

Por simplicidad, sin pérdida de generalidad, supóngase que el intercepto  $\alpha$  en la ecuación (1.1) es cero. Por definición,  $\tilde{x}_P$  maximiza con respecto a x el logaritmo de la densidad predictiva, *i.e.*, maximiza la función:

$$\ln \tilde{L}_P(x) = -\frac{1}{2} \frac{(y_0 - \hat{\beta}x)^2}{q\hat{\sigma}^2 x^2 + \hat{\sigma}^2} - \frac{1}{2} \ln(q\hat{\sigma}^2 x^2 + \hat{\sigma}^2),$$

donde  $q = \frac{1}{\sum_{i=1}^{n} x_i^2}$ .

Al tomar las derivadas de esta función con respecto a x e igualar a cero se tiene que  $\tilde{x}_P$  es una raíz del polinomio cúbico:

$$a_0 x^3 + a_1 x^2 + a_2 x + a_3, (1.30)$$

donde:

$$\begin{aligned} a_0 &= q^2 \hat{\sigma}^2, \\ a_1 &= y_0 \hat{\beta} q, \\ a_2 &= \hat{\beta}^2 + q \hat{\sigma}^2 - y_0^2 q, \\ a_3 &= -y_0 \hat{\beta}. \end{aligned}$$

Se conoce que las raíces de esta ecuación tienen la forma

$$x = \frac{z + a_1}{3a_0}$$

donde

$$z = \sqrt[3]{-\frac{N}{2} + \sqrt{\frac{N^2}{4} + \frac{M^3}{27}}} + \sqrt[3]{-\frac{N}{2} - \sqrt{\frac{N^2}{4} + \frac{M^3}{27}}},$$

y  $M = 3 (3a_0a_2 - a_1^2), N = 2a_1^3 - 9a_0a_1a_2 + 27a_0^2a_3.$ Usando las desigualdades algebraicas  $\sqrt[3]{|u|} \le (1 + |u|)$  y  $\sqrt{|u|^2 + |v|^2} \le |u| + |v|$  para cualesquiera números complejos  $u \ge v$ , donde |.| denota el módulo de números complejos, puede mostrarse que

$$|z| \le 2\left(1+|N|+\frac{|M|^{3/2}}{\sqrt{27}}\right).$$

Se debe recodar que  $y_0$  y  $\hat{\beta}$  son variables aleatorias con distribución normal y que la densidad de  $\hat{\sigma}^2$  es proporcional a una distribución chi-cuadrado. Como M y N son polinomios en variables aleatorias con distribuciones normales y chi-cuadrado, |z| tiene momentos finitos de todos los órdenes positivos. Por otra parte, al usar la desigualdad de Cauchy-Schwartz y la desigualdad algebraica  $(|u| + |v|)^2 \leq 2(|u|^2 + |v|^2)$ , puede mostrarse que

$$\mathbb{E}|\tilde{x}_P|^2 \le \sqrt{\frac{1}{3}\mathbb{E}\left(\frac{1}{a_0^4}\right)}\sqrt{4\left(\mathbb{E}|z|^4 + \mathbb{E}a_1^4\right)}.$$

Si  $\psi$  es una variable aleatoria con distribución  $\chi^2_{(d)}$  entonces

$$\mathbb{E}(\psi^r) = 2^r \frac{\Gamma(d/2+r)}{\Gamma(d/2)}, \quad \text{para todo } r \in \mathbb{R}.$$

Por tanto, como  $a_0$  tiene una densidad proporcional a una distribución chicuadrado,  $\mathbb{E}\left(\frac{1}{a_0^4}\right) < \infty$ . Esto, y el hecho de que  $\mathbb{E}|z|^4 < \infty$ ,  $\mathbb{E}a_1^4 < \infty$  implica que,

$$\mathbb{E}|\tilde{x}_P|^2 < \infty,$$

lo cual concluye la demostración.  $\square$ 

#### 1.3.4. Consistencia del estimador

En esta sección se demuestra la consistencia del estimador  $\tilde{x}_P$ .

**Teorema 1.4.** Para todo  $x_0 \in \mathbb{R}$ , se cumple que el estimador predictivo no Bayesiano  $\tilde{x}_P$  de  $x_0$  definido por (1.29) es consistente, es decir,

$$\lim_{n,m\to\infty}\tilde{x}_P =^{\mathbb{P}} x_0.$$

Demostración:

Por simplicidad, sin pérdida de generalidad, se supone que el intercepto  $\alpha$  en la ecuación (1.1) es cero. De manera análoga a como se hizo en (1.12), la densidad predictiva definida en (1.28), evaluada en  $x_0$ , para cuando existen réplicas  $y_{0i}$  =

 $\beta x_0 + e_{0i}, i = 1, ..., m$ , puede expresarse como

$$\tilde{L}_P(x_0) = \tilde{f}_P(y_0; x_0, \mathcal{D}) = \frac{1}{\sqrt{q\hat{\sigma}^2 x_0^2 + \frac{\hat{\sigma}^2}{m}}} \exp\left(-\frac{1}{2} \frac{(\overline{y}_0 - \hat{\beta} x_0)^2}{q\hat{\sigma}^2 x_0^2 + \frac{\hat{\sigma}^2}{m}}\right).$$

En la expresión anterior se tiene que:

$$\frac{1}{\sqrt{q\hat{\sigma}^2 x_0^2 + \frac{\hat{\sigma}^2}{m}}} \longrightarrow \infty \text{ cuando } m, n \to \infty$$

y que

$$\frac{(\overline{y}_0 - \hat{\beta}x_0)^2}{q\hat{\sigma}^2 x_0^2 + \frac{\hat{\sigma}^2}{m}} \sim N^2(0, 1)$$

es decir, este último factor es acotado en probabilidad; por tanto

$$\widetilde{L}_P(x_0) \xrightarrow{n,m \to \infty, \mathbb{P}} \infty.$$
(1.31)

La siguiente desigualdad algebraica puede demostrarse fácilmente: para todo u,c>0 se cumple que

$$\frac{1}{\sqrt{u}}\exp\left(-\frac{1}{2}\frac{c^2}{u}\right) \le \frac{1}{\sqrt{e}}\frac{1}{|c|}$$

Con esta desigualdad, se tiene que, para todo  $x = x_0 + \Delta \operatorname{con} \Delta \neq 0$ ,  $\tilde{L}_P(x)$  puede acotarse por

$$\widetilde{L}_P(x) \le \frac{1}{\sqrt{e}} \frac{1}{\left|\frac{\overline{y} - \widehat{\beta}(x_0 + \delta)}{\widehat{\beta}\delta}\right| |\widehat{\beta}\Delta|}$$

Como  $\overline{y} - \hat{\beta}(x_0 + \Delta) = \overline{y} - \hat{\beta}x_0 - \hat{\beta}\Delta$  tiende en probabilidad a  $\beta\Delta$ , entonces el cociente dentro del módulo tiende en probabilidad a uno, lo que implica que está acotado en probabilidad; de ahí que

$$\tilde{L}_P(x) \le \frac{1}{\sqrt{e}} \frac{1}{\left|\frac{\overline{y} - \hat{\beta}(x_0 + \Delta)}{\hat{\beta}\Delta}\right| \hat{\beta}\Delta} = \frac{1}{\sqrt{e}|\hat{\beta}\Delta|} O_{\mathbb{P}}(1)$$
(1.32)

está acotada en probabilidad (se supone, claro está, que  $\beta \neq 0$ ).

Puede verse que para todo  $\Delta_0 > 0$  se cumple que

$$P\left(\left|\tilde{x}_{P}-x_{0}\right| > \Delta_{0}\right) \leq P\left(\max_{|x-x_{0}| > \Delta_{0}} \tilde{L}_{P}(x) > \tilde{L}_{P}(x_{0})\right).$$

$$(1.33)$$

Además, (1.31) y (1.32) implican que el miembro derecho en (1.33) tiende a cero cuando m, n van a infinito. Por tanto, se tiene que

$$\lim_{n,m\to\infty} \tilde{x}_P =^{\mathbb{P}} x_0$$

lo cual completa la demostración.  $\Box$ 

## 1.4. Estudio de simulación

En la presente sección se explora la factibilidad y el comportamiento del enfoque predictivo no Bayesiano presentado para la calibración a través de un estudio de simulaciones con tamaño de muestra moderado.

El estimador predictivo  $\hat{x}_P$  se compara con otros estimadores reportados en la literatura teniendo en cuenta diferentes medidas de precisión. Además, se ilustra la utilidad del enfoque propuesto de poder explorar la verosimilitud predictiva de diferentes valores posibles del parámetro de interés  $x_0$ .

Al igual que el estimador clásico, el nuevo enfoque está principalmente orientado a usarse para problemas de calibración de extrapolación (ver comentarios en las Secciones 1.1 y 1.2). Para que un estimador sea considerado apropiado para la extrapolación, es natural imponer la condición que, hasta la aproximación de orden cero, su error cuadrático medio sea una función de  $x_0$  acotada en toda la recta real. De todos los estimadores descritos en la sección 1.1, sólo  $\hat{x}_C$ ,  $\hat{x}_{SS}$  y  $\hat{x}_N$  cumplen esta propiedad, y además son consistentes, por lo que el presente estudio se centra en dichos estimadores. El estimador inverso  $\hat{x}_I$  también se incluye, debido a su frecuente uso.

Para comparar estos estimadores, se generaron diversos conjuntos de datos para varios escenarios con diferentes valores de tamaño de muestra n y de parámetros  $\alpha$ ,  $\beta$ ,  $\sigma^2$  en el modelo (1.1), así como diferentes diseños de experimentos para seleccionar los puntos  $x_1, ..., x_n$  en el conjunto de entrenamiento  $\mathcal{D}$ . Los resultados para algunos de estos escenarios son mostrados en esta sección.

Específicamente, se pondrá  $\alpha = 0$  y  $\sigma = 0.1$  en el modelo (1.1). Nótese que, entre los tamaños de muestra finito que no cubren el análisis asintótico (dígase, n < 50), un tamaño de muestra moderado (n alrededor de 25) es notablemente interesante para muchas aplicaciones en las cuales los tamaños de muestra pequeños (dígase, n < 10) son raras veces analizados por el investigador. El tamaño de muestra se fija en este estudio en n = 24. Los puntos de diseño  $\mathcal{X} = \{x_1, ..., x_n\}$  se obtuvieron muestreando de una distribución normal con media 15 y desviación estándar 1.5. El rango verdadero de estos puntos resulta ser [12.9, 17.5], el cual será considerado el rango de calibración. Se consideraron cuatro escenarios o modelos, M1,M2, M3 y M4, correspondientes a cuatro valores diferentes de  $\beta$  en el modelo (1.1) (ver Tabla 1.1). Estos escenarios están asociados con los valores crecientes del coeficiente de variación:

$$CV = \frac{\sqrt{var(\hat{\beta})}}{|\beta|} = \frac{\sigma}{\sqrt{n}\sigma_x|\beta|}$$

Nótese que CV puede interpretarse en términos de Relación Ruido-Señal (NSR, del inglés <u>Noise to Signal Ratio</u>)  $NSR = \sigma/(\sigma_x |\beta|)$ , y es equivalente a  $\sigma^2/\beta^2$  para valores fijos de n y  $\sigma_x$ . Esta cantidad juega un papel fundamental en la precisión de los estimadores, tanto en sentido asintótico (ver sección 1.2), como en el comportamiento con tamaño de muestra finito. Sin importar el tipo de estimador usado para la calibración, escenarios con valores grandes de CV y NSR conducen a estimadores de  $x_0$ menos precisos, y se denominarán escenarios "malos" en comparación con escenarios "buenos" caracterizados por valores pequeños de CV y NSR. Puede observarse en la Tabla 1.1 que los valores de CV y NSR de esta simulación varían desde escenarios muy buenos ( $CV \approx 1\%$ ,  $NSR \approx 4\%$  en M1) hasta muy malos ( $CV \approx 50\%$ , NSR $\approx 200\%$  en M4).

	M1	M2	M3	M4
$\beta$	1.63	0.54	0.09	0.03
CV	0.009	0.029	0.177	0.531
NSR	0.044	0.142	0.867	2.601
$\sigma/eta$	0.061	0.185	1.111	3.333

Tabla 1.1: Valores de los coeficientes del modelo  $\beta$ , el coeficiente de variación CV, la razón ruido-señal NSR y  $\sigma/\beta$  para los cuatro escenarios simulados M1, M2, M3 y M4 en el estudio de simulación.

Se usaron varios criterios para evaluar el comportamiento de un estimador  $\hat{x}$  de  $x_0$ . El primero es la Raíz del Error Cuadrático Medio (RMSE, del inglés <u>Root Mean Squared Error</u>):  $RMSE = \sqrt{E(\hat{x} - x_0)^2}$ . Este criterio no puede aplicarse a estimadores que no tengan momento de segundo orden finito, tales como  $\hat{x}_C$  y  $\hat{x}_{SS}$  [104]; por tanto, se emplea sólo en las comparaciones con los estimadores  $\hat{x}_N$  y  $\hat{x}_I$  (ver sección 1.3.3 para la demostración de segundo momento finito de  $\hat{x}_P$ ). Otro criterio, adoptado como una medida de precisión, para incluir todos los estimadores  $\hat{x}_R, \hat{x}_C, \hat{x}_{SS}$  y  $\hat{x}_N$  en la evaluación es el 50% – percentil de la distribución de  $|\hat{x} - x_0|$ , denotado por D50. Para comparar estimadores que muestren un comportamiento si-

milar teniendo en cuenta RMSE y D50, se empleó la divergencia de Kullback-Leibler (KL) [55]. La divergencia KL de una estimación  $\hat{f}$  de una densidad f está definida por

$$KL(f,\hat{f}) = \int f(y)\ln f(y)dy - \int f(y)\ln \hat{f}(y)dy.$$

Para calcular y graficar las curvas RMSE y D50 como funciones de  $x_0$  se consideró un conjunto discreto de posibles valores de  $x_0$ , específicamente:

 $\mathcal{X}_0 = \{0; 2; 4; 6; 8; 10; 11; 12; 13; 14; 15; 16; 17; 18; 19; 20; 22; 24; 26; 28; 30\}.$ 

Obsérvese que  $\mathcal{X}_0$  sobrepasa el rango de calibración [12.9, 17.5], lo que permite evaluar los métodos de calibración no sólo para la interpolación (*i.e.*, dentro del rango de calibración), sino también para la extrapolación (*i.e.*, fuera del rango de calibración).

Para cada método de estimación y cada escenario M1, M2, M3 and M4, se ejecutaron los siguientes pasos computacionales:

- 1. Para cada repetición b = 1, ..., B (donde B = 2000):
  - 1a) (Generación de los datos de entrenamiento) Se simuló un conjunto de datos independientes  $\mathcal{Y}^b = (y_1^b, ..., y_n^b)$ , donde  $y_i^b$  se genera de acuerdo al modelo de regresión (1.1) con los valores de parámetros del escenario especificado y  $X = x_i$ , i = 1, ..., n. De esta forma se obtiene el conjunto de entrenamiento  $\mathcal{D}^b = \{(x_1, y_1^b), ..., (x_n, y_n^b)\}.$
  - 1b) (Generación de la observación futura) Para cada  $x_0 \in \mathcal{X}_0$ , una nueva observación independiente  $y_0^b$  se genera de acuerdo al modelo (1.1) con los valores de parámetros correspondientes al escenario especificado pero asignando  $X = x_0$ .
  - 1c) (Estimación) Para cada  $x_0 \in \mathcal{X}_0$ , los datos de entrenamiento  $\mathcal{D}^b$  y la nueva observación  $y_0^b$  se usan para obtener una estimación  $\hat{x}_0^b$  de  $x_0$ , así como la desviación asociada  $\hat{x}_0^b x_0$ .
- 2. (Cálculo de las curvas RMSE y D50) Las desviaciones para cada muestra  $\hat{x}_0^1 - x_0, ..., \hat{x}_0^B - x_0$  calculadas en el paso (1c) se usan para estimar los valores de las medidas de precisión  $RMSE(x_0)$  y  $D50(x_0)$  en cada  $x_0 \in \mathcal{X}_0$ .

Las curvas resultantes de RMSE y D50 para los diferentes métodos de calibración se grafican en la Figura 1.1 (para los escenarios M1 y M2) y la Figura 1.2 (para los escenarios M3 y M4). Ellas muestran propiedades que vale la pena destacar y que pueden resumirse como sigue:

- (I) Según ambos criterios, RMSE y D50, todos los estimadores tienen peor comportamiento a medida que  $x_0$  se aleja del centro del rango de calibración.
- (II) Como se observa en la fila superior de la Figura 1.1, en escenarios muy buenos como M1 es difícil notar diferencias entre los comportamientos de los estimadores de calibración  $\hat{x}_I$ ,  $\hat{x}_C$ ,  $\hat{x}_{SS}$ ,  $\hat{x}_N$  y  $\hat{x}_P$ , y todos ellos muestran gran precisión.
- (III) A medida que el escenario empeora, las diferencias entre los estimadores comienzan a ser notables (ver fila inferior de la Figura 1.1 y de la Figura 1.2). De manera más específica, se observa que:
  - a) Las diferencias entre el comportamiento de los estimadores  $\hat{x}_P$ ,  $\hat{x}_C$ ,  $\hat{x}_{SS}$  y  $\hat{x}_N$  son pequeñas en comparación con las diferencias entre cualquiera de ellos y el estimador inverso  $\hat{x}_I$  (ver Figura 1.1 y fila superior de la Figura 1.2) excepto para escenarios muy malos como M4 (ver fila inferior de la Figura 1.2).
  - b) En los escenarios en que alguna diferencia de comportamiento entre  $\hat{x}_I$  y el grupo de estimadores  $\hat{x}_P$ ,  $\hat{x}_C$ ,  $\hat{x}_{SS}$ ,  $\hat{x}_N$  es notable,  $\hat{x}_I$  se comporta ligeramente mejor que  $\hat{x}_P$ ,  $\hat{x}_C$ ,  $\hat{x}_{SS}$  y  $\hat{x}_N$  para valores de  $x_0$  cercanos al centro del rango de calibración (*i.e.*, para la interpolación) mientras todos los estimadores de este último grupo se comportan en gran medida, mejor que  $\hat{x}_I$  para valores de  $x_0$  fuera o cercanos a los límites del rango de calibración (*i.e.*, para la extrapolación).
  - c) A medida que el escenario empeora, como M4, (ver fila inferior de la Figura 1.2) algunas diferencias entre el comportamiento de los estimadores  $\hat{x}_P$ ,  $\hat{x}_C$ ,  $\hat{x}_{SS}$ ,  $\hat{x}_N$  se hacen notables. La curva D50 (ver esquina derecha inferior de la Figura 1.2) muestra que  $\hat{x}_P$  y  $\hat{x}_N$  tienen comportamientos similares y mejores que  $\hat{x}_C$  y  $\hat{x}_{SS}$ , que a su vez tienen un comportamiento semejante. Por tanto, para la extrapolación, se cumple el siguiente orden:  $\hat{x}_P$  y  $\hat{x}_N$  son mejores que  $\hat{x}_C$  y  $\hat{x}_{SS}$ , y estos a su vez son mejores que  $\hat{x}_I$ ; mientras para la interpolación:  $\hat{x}_I$  es mejor que  $\hat{x}_P$  y  $\hat{x}_N$  son los mejores estimadores para la extrapolación, mientras para la interpolación, proveen mejoras sobre  $\hat{x}_C$  y  $\hat{x}_{SS}$ , no así con respecto a  $\hat{x}_I$ . En escenarios marcadamente malos como M4,  $\hat{x}_N$  se comporta mejor que  $\hat{x}_P$  desde el punto de vista de RMSE.



Figura 1.1: Curvas de RMSE y D50 como función de  $x_0$  para los diferentes estimadores  $\hat{x}_I$ ,  $\hat{x}_C$ ,  $\hat{x}_{SS}$ ,  $\hat{x}_N$  y  $\hat{x}_P$  en los escenarios M1 ((a) y (b)) y M2 ((c) y (d)). Las curvas que no son visibles están superpuestas a otras curvas como las curvas correspondientes a  $\hat{x}_N$  y  $\hat{x}_P$  en (c), y las correspondientes a  $\hat{x}_C$ ,  $\hat{x}_{SS}$ ,  $\hat{x}_N$  y  $\hat{x}_P$  en (d).



Figura 1.2: Curvas de RMSE y D50 como función de  $x_0$  para los diferentes estimadores  $\hat{x}_I$ ,  $\hat{x}_C$ ,  $\hat{x}_{SS}$ ,  $\hat{x}_N$  y  $\hat{x}_P$  en los escenarios M3 ((a) y (b)) y M4 ((c) y (d)). Las curvas que no son visibles están superpuestas a otras curvas como las curvas correspondientes a  $\hat{x}_C$ ,  $\hat{x}_{SS}$ ,  $\hat{x}_N$  y  $\hat{x}_P$  en (b), y las correspondientes a  $\hat{x}_C$  y  $\hat{x}_{SS}$  en (d).

Sin embargo, esto último no significa que  $\hat{x}_N$  se comporte tan bien como  $\hat{x}_P$  respecto a otros criterios estadísticos. En particular, desde el punto de vista de la divergencia de Kullback-Liebler a la densidad verdadera  $f(y; x_0, \gamma_0, \sigma_0^2)$  de una observación futura de Y, la estimación por sustitución de la densidad  $f(y; \hat{x}_N, \hat{\gamma}, \hat{\sigma}^2)$  provista por  $\hat{x}_N$  es peor que la aproximación predictiva  $f_P(y; \hat{x}_P, \mathcal{D})$  basada en  $\hat{x}_P$ , sobre todo el rango de los posibles valores de  $x_0$ . En este sentido,  $\hat{x}_P$  conduce a una mejor estimación de la distribución de una observación futura que  $\hat{x}_N$ . Esto se ilustra en la Figura 1.3 donde, para el escenario M4, las esperanzas de las divergencias de Kullback-Liebler con respecto a  $f(y; x_0, \gamma_0, \sigma_0^2)$  correspondientes a las estimaciones de densidades por  $\hat{x}_N$  y  $\hat{x}_P$  son graficadas como función de  $x_0$ . Note que el escenario M4 es precisamente aquel en que  $\hat{x}_N$  muestra el mejor comportamiento relativo a  $\hat{x}_P$  en términos de RMSE.



Figura 1.3: Curvas de divergencia KL como función de  $x_0$  para los estimadores  $\hat{x}_N$  y  $\hat{x}_P$  en el escenario M4.

Una característica atractiva del enfoque predictivo no Bayesiano para la calibración es que provee no sólo estimaciones puntuales de  $x_0$  sino también verosimilitudes predictivas para todos los posibles valores de este parámetro. Esto nos permite complementar la estimación puntual  $\hat{x}_P$  con una apreciación basada en la verosimilitud de la localización de  $x_0$  a través de una exploración visual de la curva de verosimilitud predictiva relativa (para la importancia del análisis de la verosimilitud en la inferencia estadística, ver *e.g.* [70, 86]). A modo de ilustración, la Figura 1.4 presenta las curvas de verosimilitud predictiva relativa para muestras generadas con el escenario M1 con  $x_0 = 15$  (esquina izquierda superior), y  $x_0 = 6$  (esquina derecha superior); y muestras del escenario M3 con  $x_0 = 15$  (esquina izquierda inferior) y  $x_0 = 6$  (esquina derecha inferior). Puede observarse que las curvas de verosimilitud predictiva muestran mayor precisión en la localización de  $x_0$  (*i.e.*, son más puntiagudas) para escenarios buenos que para escenarios malos (comparar fila superior contra fila inferior en la Figura 1.4), así como para valores de  $x_0$  cerca del centro del rango de calibración que para valores lejanos (comparar columna izquierda contra columna derecha en esta figura).



Figura 1.4: Funciones de verosimilitud predictiva relativa para diferentes escenarios y valores de  $x_0$ : a)  $M1 \operatorname{con} x_0 = 15$ , b)  $M1 \operatorname{con} x_0 = 6$ , c)  $M3 \operatorname{con} x_0 = 15$  y d)  $M3 \operatorname{con} x_0 = 6$ .

Para una mayor ilustración de este punto, la Figura 1.5 muestra la función de verosimilitud predictiva para un nuevo escenario M5 en el caso extremo en que  $\beta = 0$ . Está claro que en esta situación el valor específico de  $x_0$  está indeterminado en el modelo, por tanto no tiene sentido tratar de dar alguna estimación puntual de  $x_0$ . Esto se refleja correctamente por la función de verosimilitud predictiva relativa mostrada en la Figura 1.5, la cual le asigna verosimilitudes muy altas a todos los posibles valores de  $x_0$ .



Figura 1.5: Verosimilitud predictiva relativa como función de posibles valores de  $x_0$ en un caso en que  $\beta = 0$ . El corte en 0.15 indicado por la línea horizontal discontinua corresponde al 95 % de intervalo de confianza de acuerdo con la aproximación de la chi-cuadrado al logaritmo del cociente de verosimilitudes.

# Capítulo 2

# MÁQUINAS DE VECTORES RELEVANTES PARA LA CALIBRACIÓN MULTIVARIADA

## 2.1. Introducción

Los métodos de calibración multivariada han sido los más explorados en la solución de problemas de calibración. El hecho de trabajar con varias variables en lugar de hacerlo con una sola constituyó un avance crucial, debido a la cantidad de información que se dispone en el primer caso. En el problema de la calibración multivariada se tiene que la variable Y es un vector aleatorio (típicamente mediciones obtenidas por un cierto dispositivo como un espectrómetro). En quimiometría, específicamente en espectrometría, la propiedad de interés usualmente se determina por un método de referencia que en la mayoría de los casos es costoso, laborioso y lento. Un buen modelo de calibración debe ser capaz de reemplazar el método de referencia.

Para resolver el problema de la calibración multivariada se ha propuesto una gran variedad de métodos. A continuación se presenta un resumen de los más usados, que pueden agruparse en dos grandes clases:

#### a) Métodos de calibración multivariada lineales

El estimador clásico de Eisenhart [24], que puede obtenerse siguiendo el enfoque de máxima verosimilitud bajo supuestos de normalidad, se generalizó al caso multivariado [10]. La mayoría de los estimadores propuestos son estimadores de Xlineales (en Y) que intentan aproximar el valor esperado condicional  $\mathbb{E}(X / Y)$ , como es el caso de la generalización multivariada del estimador inverso [10]. Entre los métodos más usados en este grupo, se encuentran los lineales regularizados como son la Regresión Ridge [49, 50], la Regresión por componentes principales [64], los Mínimos cuadrados parciales [39, 40, 108, 106, 107] y la Regresión continua [32, 38, 88, 89]. Su popularidad se debe a que pueden tratar la alta dimensión de Y, la alta colinealidad de sus variables, y el hecho de que usualmente la dimensión de Yexceede el tamaño n de la muestra de calibración, propiedades muy comunes en estos datos, sobre todo en la espectrometría, y que constituyen una dificultad para los otros métodos. La aplicación de estos métodos (sobre todo PLS) a la solución de problemas de calibración puede encontrarse en una inmensidad de artículos en el campo de la quimiometría. En particular, soluciones dadas a problemas concretos detectados en la industria del petróleo de nuestro país se publicaron en [20, 22].

b) Métodos de calibración multivariada no paramétricos

Una desventaja de los predictores lineales es que no siempre pueden modelar la no linealidad existente entre las variables  $X ext{ y } Y$ . Motivado por esto, varios métodos no paramétricos han sido propuestos para el problema de calibración. Las Redes neuronales son un método popular de estimación de la regresión con alta dimensión y no linealidad; sin embargo, presentan algunas desventajas [6]. En particular, su arquitectura tiene que ser determinada a priori, y además requieren estimar un número grande de parámetros no lineales, cuya determinación conduce a problemas de optimización con múltiples mínimos locales en los que resulta difícil hallar el mínimo global.

La introducción en el campo de la quimiometría de los métodos de regresión basados en vectores soportes [3, 94, 95], como son las Máquinas de vectores soportes [19, 82, 85, 100] y las Máquinas de vectores soportes de mínimos cuadrados [90] ofrecieron alternativas ventajosas sobre los enfoques existentes, pues en ellos la cantidad de parámetros a estimar no aumenta con la dimensión de Y, por tanto, no se afectan con la alta dimensión de los datos; presentan gran flexibilidad para modelar relaciones no lineales; tienen alta capacidad de generalización con pocos datos en el conjunto de calibración; a diferencia de las redes neuronales, la solución es única y global y, por último, sus soluciones son ralas, lo cual constituye una diferencia crucial con los demás enfoques. El comportamiento de estos métodos como técnica de calibración multivariada en la quimiometría ha sido estudiado por varios autores [45, 94, 95]. Como consecuencia del auge de estos métodos basados en funciones núcleos, algunos de los enfoques lineales existentes se extendieron a no lineales usando dichas funciones núcleos, como es el caso de los Mínimos cuadrados parciales con núcleos [78].

A pesar de las ventajas mencionadas, SVR y LSSVM presentan algunos inconvenientes. Por ejemplo, no permiten el uso arbitrario de funciones núcleos (*i.e.*, los núcleos deben ser definidos positivos); además, es necesario ajustar varios hiperparámetros (dos en el caso de SVR), lo cual puede ser computacionalmente costoso. Un enfoque existente en la literatura que mitiga varias de estas desventajas es el de las Máquinas de vectores relevantes para regresión, originalmente introducidas por Tipping [96, 97]. RVMR se basa en un modelo probabilístico ralo, que tiene forma funcional idéntica a SVR y LSSVM, pero adopta un enfoque de aprendizaje Bayesiano para obtener las estimaciones.

El enfoque RVMR nunca antes se había aplicado a problemas de calibración. Este es precisamente el propósito y contenido de este capítulo, en el cual se evalúa su comportamiento, en tres conjuntos de datos quimiométricos reales.

### 2.2. Máquinas de vectores relevantes para regresión

Tipping [96] propuso las Máquinas de vectores relevantes para retomar las ideas principales de SVR en un contexto Bayesiano. Dado un conjunto de datos de entrenamiento  $\mathcal{D} = \{x_i, \mathbf{y}_i\}_{i=1}^n$ , el siguiente modelo de regresión lineal generalizado se usa para describir la relación de mapeo entre el vector de patrones de entrada  $\mathbf{y}$ , y el escalar x:

$$x_i = g(\mathbf{y}_i; \mathbf{w}) + \epsilon_i,$$

donde los errores  $\boldsymbol{\epsilon} = (\epsilon_1, ..., \epsilon_n)$  se suponen independientes, Gaussianos con media cero y varianza  $\sigma^2$ ;  $\mathbf{w} = (w_1, ..., w_m)$  es el vector de parámetros del modelo y  $g(\mathbf{y}_i; \mathbf{w})$ se expresa como una suma lineal ponderada de funciones bases  $\phi_j(\mathbf{y})$  especificadas

$$g(\mathbf{y}, \mathbf{w}) = \sum_{j=1}^{m} w_j \phi_j(\mathbf{y}) + w_0, \qquad \mathbf{x} = \Phi \mathbf{w}.$$
 (2.1)

Aquí  $\Phi = [\phi_1, ..., \phi_m]$  es la matriz de diseño  $n \times m$  cuyas columnas contienen el conjunto completo de los m vectores bases. La forma de la función (2.1) es igual a la forma que tiene la función del método SVR si las funciones bases se toman como las funciones núcleos en SVR parametrizadas con los datos de calibración:  $\phi_j(\mathbf{y}) = K(\mathbf{y}, \mathbf{y}_j)$  y  $\phi(\mathbf{y}_i) = [1, K(\mathbf{y}_1, \mathbf{y}_i), ..., K(\mathbf{y}_n, \mathbf{y}_i)].$  El modelo de error supuesto implica que  $p(x_i \neq y_i) = \mathcal{N}(g(\mathbf{y}_i), \sigma^2)$ , donde la notación especifica que  $x_i$  se distribuye como una Gaussiana con media  $g(\mathbf{y}_i)$  y varianza  $\sigma^2$ . Debido a los supuestos de independencia de los  $x_i$ , la verosimilitud de la muestra de entrenamiento puede escribirse:

$$p(\mathbf{x}/\mathbf{w},\sigma^2) = (2\pi\sigma^2)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2} \| \mathbf{x} - \Phi \mathbf{w} \|^2\right\}.$$

Las estimaciones máximo verosímiles de  $\mathbf{w}$  y  $\sigma^2$  a partir de la ecuación anterior generalmente conducen a problemas de sobreajuste. RVMR, en cambio, adopta una perspectiva Bayesiana, y restringe los valores de los parámetros definiendo una distribución de probabilidad a priori para  $\mathbf{w}$ , que expresa preferencia por valores pequeños de  $\mathbf{w}$  a través de la elección usual de distribuciones a priori Gaussianas con media cero y varianza  $\alpha_i^{-1}$ :

$$p(\mathbf{w} \nearrow \boldsymbol{\alpha}) = (2\pi)^{-m/2} \prod_{j=1}^{m} \alpha_j^{1/2} \exp\left(-\frac{\alpha_j w_j^2}{2}\right)$$

con  $\boldsymbol{\alpha} = (\alpha_1, ..., \alpha_m)$  un vector de *m* hiperparámetros. Debe destacarse que hay un hiperparámetro individual asociado de manera independiente a cada ponderación. Esta forma de la distribución a priori es la responsable de la raleza del método [97].

Conocido  $\alpha$  (de madera análoga puede hacerse para  $\sigma$ ), la distribución a posteriori de los parámetros puede obtenerse combinando la verosimilitud y la distribución a priori según la regla de Bayes:

$$p(\mathbf{w}/\mathbf{x}, \boldsymbol{\alpha}, \sigma^2) = \frac{p(\mathbf{x}/\mathbf{w}, \sigma^2)p(\mathbf{w}/\boldsymbol{\alpha})}{p(\mathbf{x}/\boldsymbol{\alpha}, \sigma^2)},$$
(2.2)

la cual resulta ser una distribución Gaussiana  $\mathcal{N}(\boldsymbol{\mu}, \Sigma)$  con

$$\boldsymbol{\mu} = \sigma^{-2} \Sigma \Phi^T \mathbf{x}, \qquad \Sigma = (\mathbf{A} + \sigma^{-2} \Phi^T \Phi)^{-1}$$
(2.3)

y  $\mathbf{A} = diag(\alpha_1, ..., \alpha_M)$ . En vez de continuar el desarrollo y hacer inferencia Bayesiana sobre estos hiperparámetros (lo cual conduce a integrales que no tienen expresión analítica exacta), el aprendizaje Bayesiano ralo (<u>Sparse Bayesian learning</u>) [97] puede formularse como un procedimiento *de máxima verosimilitud tipo II*; esto es: se encuentra la estimación puntual más probable  $\boldsymbol{\alpha}_{\text{MP}}$  maximizando la *verosimilitud marginal*, o equivalentemente, su logaritmo  $\mathcal{L}(\boldsymbol{\alpha})$ , con respecto al hiperparámetro  $\boldsymbol{\alpha}$ :

$$\mathcal{L}(\boldsymbol{\alpha}) = \log p(\mathbf{x}/\boldsymbol{\alpha}, \sigma^2) = \log \int_{-\infty}^{\infty} p(\mathbf{x}/\mathbf{w}, \sigma^2) p(\mathbf{w}/\boldsymbol{\alpha}) d\mathbf{w}$$
$$= -\frac{1}{2} [N \log 2\pi + \log |\mathbf{C}| + \mathbf{x}^{\mathrm{T}} \mathbf{C}^{-1} \mathbf{x}],$$

donde

$$\mathbf{C} = \sigma^2 \mathbf{I} + \Phi \mathbf{A}^{-1} \Phi^{\mathrm{T}}$$

Una vez que se determinan los valores más probables  $\alpha_{\rm MP}$  y  $\sigma_{\rm MP}$ , se obtiene una estimación puntual  $\mu_{\rm MP}$  para los parámetros evaluando (2.3) con  $\alpha = \alpha_{\rm MP}$  y  $\sigma = \sigma_{MP}$ . La observación importante es que típicamente los valores óptimos para muchos de los hiperparámetros  $\alpha_j$  son infinitos [97]. De (2.3), esto conduce a una distribución a posteriori de los parámetros  $w_m$  infinitamente puntiaguda en cero, con la consecuencia de que  $\mu_{\rm MP}$  contendrá pocos elementos distintos de cero. Aquellos vectores soportes asociados a las ponderaciones  $w_m$  diferentes de cero se denominan relevantes, en referencia al principio de determinación automática de relevancia (del inglés, automatic relevance determination)[97].

Las predicciones se hacen sobre la base de la distribución de las ponderaciones condicionadas por los valores máximos  $\alpha_{\rm MP}$  y  $\sigma_{\rm MP}^2$ . Específicamente, la distribución predictiva para un nuevo dato  $\mathbf{x_0}$ , usando (2.2) queda definida como

$$p(x_0 \not \mathbf{x}, \boldsymbol{\alpha}_{\mathrm{MP}}, \sigma_{\mathrm{MP}}^2) = \int p(x_0 \not \mathbf{w}, \sigma_{MP}^2) p(\mathbf{w} \not \mathbf{x}, \boldsymbol{\alpha}_{\mathrm{MP}}, \sigma_{\mathrm{MP}}^2) d\mathbf{w},$$

la cual puede calcularse fácilmente debido al hecho de que ambos términos en el integrando son Gaussianos, por tanto el resultado es también una Gaussiana  $\mathcal{N}(\mu_0, \sigma_0^2)$ :

$$\mu_0 = \boldsymbol{\mu}^T \phi(\mathbf{y}_0), \qquad \sigma_0^2 = \sigma_{MP}^2 + \phi(\mathbf{y}_0)^T \Sigma \phi(\mathbf{y}_0).$$

Puede verse que, de manera intuitiva, el predictor de la media  $\mu_0$  es la función (2.1) evaluada en la media a posteriori de los pesos, es decir  $g(\mathbf{y}_0; \boldsymbol{\mu})$ ; obteniéndose como estimador final  $\mathbf{x} = \Phi \boldsymbol{\mu}_{\text{MP}}$ .

### 2.3. Aplicación a datos reales

El método RVMR se aplica a tres conjuntos de datos reales. Los resultados son comparados con los obtenidos usando otros métodos de calibración como PLS, SVR y LSSVM, cuyos comportamientos fueron estudiados previamente en la literatura [95].

#### 2.3.1. Conjunto de datos "Wülfert"

Este conjunto de datos está constituido por espectros de Infrarrojo Cercano (NIR, del inglés <u>Near Infrared</u>) de mezclas de etanol, agua y 2-propanol, originalmente medidas y descritas por Wülfert y col. [109]. Wülfert propuso un diseño de experimentos de mezclas formadas por 19 combinaciones diferentes de fracción molar de los tres constituyentes, las cuales fueron analizadas en el rango de longitudes de onda de 850 - 1049 nm con resolución de 1 nm (200 longitudes de onda). Cada una de las 19 mezclas se midió a cinco temperaturas diferentes: 30, 40, 50, 60, 70°C. El objetivo con estos datos es predecir, a partir de la información espectral, la fracción molar de etanol, agua y 2-propanol en una mezcla. El reto está en que los cambios de temperatura causan variaciones espectrales no lineales, por tanto, las relaciones entre los espectros medidos a diferentes temperaturas no pueden determinarse de manera lineal.

En estudios previos [95] se separaron las 19 combinaciones de fracción molar en 6 mezclas específicas para conformar el conjunto de prueba y las 13 restantes para calibrar. Para poder comparar los resultados de esta experimentación con los existentes en la literatura, se realizó esta división de la misma forma.

Estos datos se trataron desde dos puntos de vista teniendo en cuenta los cambios de temperatura: local y global. Desde el punto de vista local, se tienen cinco conjuntos de calibración con sus respectivos conjuntos de prueba (uno por cada temperatura) con tamaños muestrales de 13 y 6, respectivamente. De esta forma, tanto la construcción de los modelos de calibración, como las predicciones, se hacen para cada temperatura por separado. En cambio, desde el punto de vista global, todas las mezclas se unen y se tiene un único conjunto de calibración de tamaño muestral 65 y un conjunto de prueba de tamaño muestral 30. En los experimentos que se presentan a continuación todos los modelos de calibración construidos usando RVMR son globales, de esta forma se enfrenta el reto de la no linealidad debido al factor temperatura.

Los datos se preprocesaron de la misma forma que indica el trabajo de Wülfert [109], *i.e.*, corrección de la línea base y centrado por la media.

#### Resultados y comparación con otros enfoques

A modo de verificación, se reprodujeron los resultados reportados en la literatura alcanzados por los métodos SVR (modelo global) [94] y LSSVM (modelo global) [95]. Los resultados de PLS seleccionados fueron los correspondientes a los modelos locales [94] porque los modelos globales presentan un comportamiento pobre (RMSE reportado de 0.0195 para el etanol, 0.085 para el agua y 0.023 para el 2-propanol).

Para la construcción del modelo RVMR se empleó un núcleo Gaussiano y su parámetro ancho de banda h se ajustó haciendo una validación cruzada con 5 particiones en el conjunto de entrenamiento. Para esto, se exploraron diferentes rangos de posibles valores de h por cada compuesto: 0.01 - 0.65, 0.01 - 1 y 0.01 - 0.6 (con paso 0.05) para el etanol, el agua y el 2-propanol, respectivamente.



Figura 2.1: Valores de RMSE durante la validación cruzada (RMSECV) para (a)etanol (b)agua (c)2-propanol. El valor óptimo de  $\sigma$  aparece indicado con un asterisco.

Los valores de h que se obtuvieron, como muestra la Figura 2.1, son h = 0.45, h = 0.25 y h = 0.55 para el etanol, el agua y el 2-propanol, respectivamente.

La Figura 2.2 compara los errores de predicción, usando la Raíz del Error Cuadrático Medio de Predicción (RMSEP, del inglés <u>Root Mean Square Error of Prediction</u>), para los diferentes métodos.



Figura 2.2: Valor de RMSEP alcanzado por cada método en el conjunto de prueba para cada uno de los compuestos

El método RVMR, al igual que los otros métodos no lineales, se comporta mejor que el método PLS. LSSVM es ligeramente más preciso que RVMR (en un factor de 1.16), pero este último tiene mejor comportamiento que SVR (en un factor de 1.08), excepto para el caso del agua.

La Figura 2.3 muestra una representación gráfica de los valores reales contra los valores predichos por el método RVMR para cada uno de los compuestos en el conjunto de prueba. Esto confirma que la exactitud de las predicciones son bastante buenas para todos los compuestos.



Figura 2.3: Valores reales contra valores predichos para (a)Etanol (b)Agua y (c)2-propanol por el método RVMR

La Tabla 2.1 presenta el número de vectores soportes para SVR y LSSVM [94, 95] y el número de vectores relevantes para RVMR, necesarios para la predicción de los tres compuestos químicos en mezclas futuras.

	Etanol		Agua		2-propanol	
	# sv	%	$\# \mathbf{sv}$	%	$\# \mathbf{sv}$	%
SVR	39	60	27	41.5	36	55
LSSVM	37	57	57	88	35	54
	$\# \mathbf{rv}$	%	$\# \mathbf{rv}$	%	$\# \mathbf{rv}$	%
RVMR	13	20	16	24	14	21

Tabla 2.1: Número de vectores soportes (sv) y vectores relevantes (rv), y porciento que representan del total de la muestra de calibración, para los distintos métodos.

Es evidente que RVMR es el método con mayor raleza. En la mayoría de los casos se obtiene una raleza de más de la mitad con respecto a los otros métodos.

LSSVM en su versión original usa todos los datos del conjunto de calibración en el modelo estimado. Para alcanzar estimaciones del modelo ralas es necesario aplicar técnicas de poda a los multiplicadores de Lagrange [90] (*i.e.*, se ponen como cero aquellos cuyos valores no sobrepasan un umbral especificado), lo cual implica un incremento tanto del error de predicción como del costo computacional.

Para el caso del agua, RVMR no logra superar los otros métodos no lineales. Cuando Thissen y col. [95] podaron el modelo LSSVM estimado correspondiente al agua, no pudieron bajar el número de vectores soportes hasta el alcanzado por SVR (41.5%) porque el error de predicción aumentó a 0.0071. Para alcanzar resultados comparables con SVR, ellos tuvieron que mantener el número de vectores de soporte igual a 57 (88%). Sin embargo, RVMR alcanza en este caso casi la mitad del número de vectores usados en SVR. Es posible que estos niveles de raleza tan altos alcanzados por RVMR atenten contra la precisión en este caso.

Otro aspecto de interés es mostrar la influencia de cada dato de calibración en la solución final. Para tales propósitos se utilizaron los valores estimados de varianza inversa ( $\alpha_i^{-1}$ ). En RVMR, la estimación del parámetro de varianza inversa  $\alpha_i^{-1}$  puede ser considerado como expresión de la "importancia" o la "relevancia" de cada dato: los valores más altos de  $\alpha_i$  (*i.e.*, los menores de  $\alpha_i^{-1}$ ) son asignados a los vectores menos relevantes para el modelo.

La Figura 2.4 muestra los datos de calibración más importantes para los modelos estimados asociados a cada compuesto (vectores relevantes), así como aquellos que no contribuyen a la solución ( $\alpha_i^{-1} \cong 0$ ).



Figura 2.4: Valor de varianza inversa correspondiente a cada datos de calibración para los modelos de (a)Etanol (b)Agua y (c)2-propanol

Debido a que los valores de varianza inversa difieren en magnitud para cada modelo (nótese que las escalas son diferentes) se decidió estudiar la contribución de los datos en cada modelo estimado por separado. Para mostrar la localización de los datos más importantes en el diseño de las mezclas propuesto por Wülfert y col., se siguió el procedimiento explicado por Thissen [95]. La importancia de cada punto de mezcla en el diseño se obtuvo tomando el promedio de los 5 diseños de mezclas individuales correspondientes a los cinco valores de temperatura (ver Figura 2.5).



Figura 2.5: Importancia relativa de los datos de calibración para los modelos de (a)Etanol (b)Agua (c)2-propanol. La importancia relativa fue representada en una escala de grises: mientras más oscuro es el color, más importante es el dato para el modelo.

Los diseños de mezclas correspondientes a los modelos estimados para predecir cada uno de los compuestos muestran diferentes distribuciones de los datos más importantes. También es diferente la importancia relativa de cada dato en cada diseño de mezclas. Sin embargo, hay una tendencia común en todos los diseños. Puede verse que todos ellos usan datos de calibración con fracción molar alta de etanol y 2-propanol y fracción molar baja de agua. Thissen y col.[95] obtuvieron comportamientos similares al usar el método SVR, los cuales fueron argumentados por el hecho que los espectros NIR de etanol y 2-propanol son similares mientras el espectro del agua es mucho más diferente, por tanto es más difícil distinguir etanol de 2-propanol que etanol de agua.

### 2.3.2. Conjunto de datos "Tecator"

El conjunto de datos denominado Tecator proviene de la industria alimenticia [92] y contiene 215 espectros NIR de muestras de carne, medidas en un equipo <u>Tecator Infratec Food and Feed Analyzer</u>. Cada observación consta de 100 valores de absorbancia en el rango de longitudes de onda 850 – 1050 nm y está asociada a una descripción de la muestra de carne, obtenida por un químico analítico que contiene el porciento de grasa, agua y proteína de cada muestra. El problema de calibración abordado con este conjunto de datos consiste en la predicción del porcentaje de grasa de la muestra de carne a partir de su espectro. De los 215 espectros, 43 se mantienen aparte como conjunto de prueba y las 172 muestras restantes se usan para la estimación del modelo (conjunto de calibración). Los espectros originales se procesan para tener media cero y varianza unitaria. Este es un conjunto de datos donde se conoce también que la relación entre la propiedad de interés y las mediciones espectrales es no lineal [7].

#### Resultados y comparación con otros enfoques

En este conjunto de datos, se usaron los métodos SVR y LSSVM para comparar su comportamiento con el del nuevo método de calibración propuesto, RVMR. Los resultados obtenidos por PLS se tomaron de la literatura.

Tanto SVR como LSSVM se aplicaron usando un núcleo Gaussiano. Para seleccionar los hiperparámetros del modelo se exploró una rejilla de búsqueda basada en una validación cruzada con 4 particiones, lo que dio como resultados los valores  $C = 1000, \sigma = 0.97, \epsilon = 0.5$  para SVR y  $\gamma = 2989.21, \sigma = 2.13$  para LSSVM. En el caso de RVMR, se usó un núcleo spline lineal y su parámetro de escala seleccionado fue  $\eta = 3.52$ .

La Figura 2.6 muestra la comparación de los errores de predicción del contenido de grasa para los diferentes métodos.



Figura 2.6: Valor de RMSEP alcanzado por cada método en el conjunto de prueba

En la Figura 2.6 se aprecia que todos los métodos no lineales se comportan mejor que el tradicionalmente usado PLS y que RVMR es ligeramente más preciso que LSSVM. El gráfico de los valores reales contra los valores predichos en el conjunto de prueba por el modelo RVMR se muestra en la Figura 2.7.



Figura 2.7: Valores reales contra valores predichos para el método RVMR.

La Tabla 2.2 confirma que RVMR es drásticamente más ralo que SVR y LSSVM. El método RVMR requiere solo el 20% del conjunto de calibración para futuras predicciones, menos de un cuarto de las muestras de calibración requeridas por los otros métodos.

	SVR	LSSVM	RVMR
$\# \mathbf{sv/rv}$	147	148	35
%	85	86	20

Tabla 2.2: Número de vectores soportes (vectores relevantes) y porciento que representan de la muestra de calibración, para los distintos métodos

Al igual que en la experimentación con el conjunto de datos anterior, resulta importante explorar la influencia de cada dato del conjunto de calibración en los modelos estimados correspondientes a SVR, LSSVM y RVMR. Los valores de varianza inversa estimados se tuvieron en cuenta para el modelo correspondiente a RVMR y los multiplicadores de Lagrange para los modelos correspondientes a SVR y LSSVM. Nótese que existen datos que son usados por los tres métodos, así como otros que no son usados por ninguno de los tres. Estos resultados se muestran en la Figura 2.8.



Figura 2.8: Importancia de los datos de calibración para los modelos estimados correspondientes a (a)SVR (b)LSSVM and (c)RVMR.

### 2.3.3. Conjunto de datos "Tablet"

Este conjunto de datos está formado por tabletas farmacéuticas de un tipo de medicamento con variabilidad en la concentración de su componente activo. Se cuenta con 120 espectros infrarrojos de tramitancia medidos en el intervalo de longitudes de onda de 7400 cm<sup>-1</sup> a 10507 cm<sup>-1</sup>, para un total de 404 variables, correspondientes a la producción de escala piloto del producto [23, 91].

El problema de interés es predecir el contenido de compuesto activo (w/w) de una tableta farmacéutica a partir de su caracterización espectral. El conjunto de datos completo fue dividido en 65 muestras para la calibración y 55 muestras para el conjunto de prueba. Los espectros fueron preprocesados usando la técnica de Corrección Multiplicativa de la Señal (MSC, del inglés <u>Multiplicative Scatter Correction</u>) [33]. Se conoce de la literatura [23] que la estructura inherente de estos datos es aproximadamente lineal y que PLS ofrece predicciones satisfactorias, de modo que este conjunto de datos nos permite investigar si el método de calibración propuesto se comporta bien en problemas lineales.

#### Resultados y comparación con otros enfoques

En este conjunto de datos, los métodos SVR y LSSVM se usaron para comparar su comportamiento con el del nuevo método de calibración propuesto, RVMR. Los resultados que se muestran correspondientes a PLS se reprodujeron de la literatura [23]. Tanto para SVR, LSSVM como para RVMR se utilizó un núcleo Gaussiano. Para la selección de los hiperparámetros se exploró una rejilla de búsqueda basada en una validación cruzada con 10 particiones. Los valores de los hiperparámetros resultantes fueron: C = 10,  $\epsilon = 0.2$ ,  $\sigma = 0.61$  para SVR;  $\gamma = 46.261$ ,  $\sigma = 0.65$  para LSSVM; y ancho de núcleo  $\eta = 0.7$  para RVMR.

La Figura 2.9 muestra los errores de predicción de contenido activo en tabletas para los diferentes métodos.



Figura 2.9: Valor de RMSEP alcanzado por cada método en el conjunto de prueba

Como es conocido, el método PLS se comporta razonablemente bien en este conjunto de datos, lo que corrobora la relación lineal entre los espectros y la propiedad de interés. No obstante, los métodos no lineales resultan más precisos, aunque en este caso las diferencias con respecto al método PLS no son tan significativas como en los casos anteriores presentados. RVMR es ligeramente más preciso que SVR y LSSVM. Tanto RVMR, como los otros métodos no lineales, han demostrado comportarse bien en la solución de problemas lineales. A pesar de la alta flexibilidad del método, es posible estimar una función de regresión simple sin sobreajustar los datos de calibración. La Figura 2.10 corrobora la exactitud del método RVMR.

El mismo número de vectores soportes fue seleccionado por los métodos SVR y LSSVM. Nuevamente, RVMR es mucho más ralo, reteniendo sólo un 6% de los datos de calibración. Estos resultados se muestran en la Tabla 2.3

La influencia de los datos del conjunto de calibración en cada uno de los modelos estimados se presenta en la Figura 2.11. Aquellos datos que no influyen en ninguno de



Figura 2.10: Valores reales contra valores predichos de w/w para el modelo RVMR

	SVR	LSSVM	RVMR
$\# \mathbf{sv/rv}$	34	34	4
%	52	52	6

Tabla 2.3: Número de vectores soportes (vectores relevantes) y porciento que representan de la muestra de calibración, para los distintos métodos.

los modelos fueron marcados en negro. Obsérvese que todos los individuos usados por RVMR fueron usados también por los otros métodos, SVR y LSSVM. Esto sugiere que RVMR, en su estrategia de selección de los datos más importantes, no incluye datos nuevos que pudieran pensarse "importantes" bajo otros criterios distintos a los que sigue SVR y SLSVM; sino que se queda con un subconjunto de los datos considerados importantes por los otros métodos.



Figura 2.11: Importancia de los datos de calibración para los modelos estimados correspondientes a (a)SVR (b)LSSVM y (c)RVMR

Debe resaltarse que la predicción de la propiedad de interés en datos futuros usando RVMR es mucho menos costoso computacionalmente que cuando se usan los otros métodos debido a la raleza de RVMR. Esto constituye una ventaja importante si las predicciones de las propiedades de interés se necesitan hacer en tiempo real (<u>on line</u>).

# Capítulo 3

# MÁQUINAS DE VECTORES SOPORTES PARA DATOS FUNCIONALES

# 3.1. Revisión de los enfoques previos de regresión funcional

El rápido desarrollo de los equipos de análisis instrumental y otros dispositivos de medición modernos proveen gran cantidad de datos como funciones digitalizadas de alta resolución. Como consecuencia, el Análisis de Datos Funcionales se ha convertido en un campo de reciente investigación [27, 76]. En el contexto del FDA, cada individuo es tratado como una función continua que toma valores en los números reales, en vez de por un vector de dimensión finita. En el problema de calibración para datos funcionales, entonces, se observa una variable aleatoria funcional Y (típicamente con valores en un espacio de Hilbert  $\mathcal{Y}$  con producto escalar  $\langle \cdot, \cdot \rangle$ ).

Hasta la actualidad, los métodos empleados para el problema de calibración con datos funcionales han estado orientados a aproximar la función de regresión  $\gamma(y) = \mathbb{E}(X \swarrow Y = y)$ , *i.e.*, a construir estimadores que minimicen el riesgo Bayesiano cuadrático. Equivalentemente, el objetivo es ajustar el modelo de regresión funcional

$$X = \gamma(Y) + \epsilon,$$

donde  $\epsilon$  es una variable aleatoria independiente de Y.

Para dar solución a este problema se han propuesto varios enfoques, los cuales

pueden ser clasificados en tres familias fundamentales:

a) Métodos de regresión funcional lineal. Los primeros trabajos estuvieron enfocados a modelos de regresión lineal donde la función de regresión  $\gamma$  tiene la forma lineal

$$\gamma(y) = c + \langle \beta, y \rangle,$$

donde  $c \in \mathbb{R}$  y  $\beta \in \mathcal{Y}$  son parámetros desconocidos. Una revisión de los enfoques para estimar este modelo puede encontrarse en [76]; ver también, [12, 13, 15, 18, 38, 62, 74].

b) Métodos de regresión funcional no paramétricos. Por su naturaleza, los métodos de regresión lineales no pueden tratar con dependencias no lineales entre las variables predictora y respuesta. Para vencer esta dificultad, se ha propuesto un número de enfoques no paramétricos. Seguidamente, se describen los enfoques más importantes desarrollados en los últimos años en esta área.

El primer enfoque, introducido por Ferraty y Vieu [27], es el uso de los estimadores de regresión por núcleos funcionales:

$$\widehat{\gamma}(y) = \frac{\sum_{i=1}^{n} K(d(y_i, y)/h)y_i}{\sum_{i=1}^{n} K(d(y_i, y)/h)},$$

donde h > 0 es el ancho de banda del núcleo, d es una semi-métrica en  $\mathcal{Y}$  y K:  $\mathbb{R}_+ \to \mathbb{R}_+$  es una función núcleo apropiada. Este tipo de estimadores permite una gran flexibilidad para ajustar modelos no lineales. Sin embargo, sus propiedades de consistencia se han demostrado sólo para clases restrictivas de funciones núcleos, como son las tipos I y II definidas en [27]; además, la selección del parámtro ancho de núcleo h sobre la base de los datos de calibración es un problema difícil, especialmente para casos como estos de alta dimensión, por lo que continúa siendo tema de investigación.

Otra clase de estimadores de regresión no paramétricos son las redes neuronales funcionales propuestas por Rossi y col. [79]. En particular, el perceptrón de una sola capa se define por

$$\widehat{\gamma}(y) = \sum_{j=1}^{q} \widehat{a}_j T(\widehat{u}_j + \widehat{l}_j(y)),$$

donde  $T : \mathbb{R} \to \mathbb{R}$  es una función de activación dada, las  $(l_j)_j$  son funcionales lineales a ser estimadas  $(e.g., l_j(y) = \langle w_j, y \rangle \text{ con } w_j \in \mathcal{Y})$  y  $(a_j)_j, (u_j)_j \in \mathbb{R}^q$  son parámetros desconocidos que también deben estimarse. El perceptrón funcional tiene la propiedad de aproximación universal que hace posible representar una gran variedad de funcionales no lineales. Pero nótese que dependen de un número de parámetros bastante grande  $(w_j)_j$ ,  $(a_j)_j$ ,  $(u_j)_j$ , los cuales aumentan con el número de neuronas q, y su estimación por mínimos cuadrados conduce a problemas de mínimos locales. Además, tiene que seleccionarse el número de neuronas, lo cual es una tarea computacionalmente difícil. En el mismo espíritu, se introdujeron versiones funcionales de las redes neuronales de funciones de bases radiales [80].

El marco general de aproximación de funciones en espacios de Hilbert con núcleos reproductores se ha usado también para introducir estimadores de regresión funcionales [73], los cuales tienen la forma general:

$$\widehat{\gamma}(y) = \sum_{i=1}^{n} \widehat{a}_i K(y_i, y)$$

donde  $K : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$  es un núcleo reproductor en  $\mathcal{Y}$ , y  $(a_i)_i \in \mathbb{R}$  son parámetros desconocidos. En este marco se han propuesto versiones funcionales de aproximación por funciones de bases radiales [73], que tienen la forma:

$$\widehat{\gamma}(y) = \sum_{i=1}^{m} \widehat{a}_i \phi(d(y, c_i)),$$

donde  $\phi : \mathbb{R}_+ \to \mathbb{R}$  es la función de base radial adoptada,  $c_1, ..., c_m \in \mathcal{Y}$  son centros dados, d es una distancia definida en  $\mathcal{Y}$ , y  $(a_i)_i$  son parámetros desconocidos.

Una ventaja importante del enfoque de aproximación de funciones en RKHS es que el estimador resultante es lineal con respecto a los parámetros desconocidos  $(a_i)_i$ . Por tanto, su estimación por optimización de mínimos cuadrados se reduce a resolver un problema lineal algebraico. Sin embargo, de manera contraria a los métodos estándares en RKHS para aproximar funciones multivariadas (*e.g.*, splines multivariados estándar y funciones de base radial como <u>thin-plate splines</u>), en el contexto del FDA las propiedades de suavidad de  $\hat{\gamma}$  como un funcional en  $\mathcal{Y}$  no han sido definidas todavía. Por tanto, la selección de núcleos reproductores y funciones de base radiales apropiados, permanece aún como un problema abierto de investigación.

Para finalizar, la regresión usando k-vecinos más cercanos ha sido generalizada a

datos funcionales [5, 56]. Este enfoque conduce a la siguiente función de regresión:

$$\hat{\gamma}(y) = \frac{1}{k} \sum_{i=1}^{k} x_{(i,y)}$$

donde  $x_{(i,y)}$  es el valor de X para el *i-ésimo* vecino más cercano de y dentro de la muestra  $(y_i)_{i=1,...,n}$ . La consistencia y la razón de convergencia de este enfoque se han estudiado suponiendo ciertas condiciones de regularidad en  $\gamma$ . Este estimador es similar al estimador de núcleos, pero tiene la limitación que brinda soluciones menos suaves, especialmente cuando el número de datos en la muestra de calibración es pequeño.

c) Métodos de Regresión Inversa Funcional. Recientemente, se propuso un enfoque alternativo que puede verse como un compromiso entre los métodos paramétricos muy restrictivos (e.g., la regresión lineal) y los no paramétricos (e.g., métodos basados en núcleos). Este se denomina Regresión Inversa Funcional [21, 29, 30], y constituye una versión funcional de la Regresión Inversa Partida previamente propuesta para modelos de regresión multivariados [57]. El enfoque FIR supone que el siguiente modelo se cumple:

$$X = g(\langle \beta_1, Y \rangle, ..., \langle \beta_d, Y \rangle) + \epsilon,$$

donde d es la llamada dimensión efectiva y  $g : \mathbb{R}^d \to \mathbb{R}$  es una función desconocida. Bajo algunos supuestos adicionales (los cuales están garantizados si Y tiene distribución elíptica, *e.g.*: una distribución Gaussiana en el espacio de Hilbert  $\mathcal{Y}$ ) las direcciones  $(\beta_j)_j$  pueden ser estimadas de la descomposición espectral del operador de covarianza  $\mathbb{V}(Y)$  y  $\mathbb{V}(\mathbb{E}(Y \nearrow X))$ . Esto último involucra ajustar la media del modelo inverso

$$Y = \mu(X) + e, \tag{3.1}$$

donde e es un elemento aleatorio en  $\mathcal{Y}$  con media cero, no correlacionado con X. De manera más específica, para estimar  $(\beta_j)_j$  y g, se llevan a cabo los siguientes pasos:

- 1. Obtención de un estimador de  $\mu(X) = \mathbb{E}(Y \nearrow X)$  haciendo la regresión de cada componente de Y en X a través de cualquier método de regresión no paramétrico univariado.
- 2. Obtención de las estimaciones de los operadores de covarianza  $\Gamma = \mathbb{V}(Y)$  y  $\Gamma_e =$

 $\mathbb{V}(\mathbb{E}(Y \nearrow X))$  sobre la base de los resultados del paso (1), y de la descomposición espectral del operador  $\Gamma^{-1/2}\Gamma_e\Gamma^{-1/2}$ . Los  $(\beta_j)_j$  se estiman como las funciones propias correspondientes a los d primeros (mayores) valores propios.

3. Estimación de g a través de un método de regresión no paramétrico d-variado.

En este enfoque semi-paramétrico, la dimensión d es un hiperparámetro del modelo. Se han propuesto varios métodos para encontrar un buen valor para d, como el referido en [28].

En el presente capítulo se desarrolla un nuevo enfoque para la estimación funcional no paramétrica que, a diferencia de los enfoques no paramétricos previamente descritos, conduce a la reducción de la dimensión; es decir, es ralo. Además, no tiene las limitaciones distribucionales de FIR, que es el único enfoque existente que contempla reducción de la dimensión. Específicamente, se propone una extensión a datos funcionales del método Máquinas de vectores soportes que se denominará Máquinas de Vectores Soportes Funcionales (FSVR). El método SVR se ha explorado mucho en el contexto de la calibración multivariada, y muestra buenos resultados (como pudo apreciarse en el Capítulo 2), por lo que su extensión a datos funcionales resulta promisoria.

# 3.2. Conceptos preliminares sobre la regresión no paramétrica por regularización en espacios de Hilbert con núcleos reproductores

Varios autores han elaborado diferentes métodos de estimación para modelos de regresión sobre la base de la regularización en espacios de Hilbert con núcleos reproductores (ver por ejemplo [9, 25, 82] y sus referencias, para definiciones y detalles). Aquí se describe brevemente este marco abstracto antes de considerar su especificación para la regresión con datos funcionales en la próxima sección.

Sea  $\mathcal{Y}$  un espacio vectorial con norma  $\|\cdot\|_{\mathcal{Y}}$ , y sea  $\mathbb{R}^{\mathcal{Y}}$  el conjunto de funciones de  $\mathcal{Y}$ en  $\mathbb{R}$ . Supóngase dada alguna función definida positiva (o núcleo)  $\kappa : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ . Es conocido que existe un RKHS  $\mathcal{H}_{\kappa} \subset \mathbb{R}^{\mathcal{Y}}$  con núcleo reproductor  $\kappa$  [82]. Se denota por  $\|\cdot\|_{\mathcal{H}_{\kappa}}$  la norma en  $\mathcal{H}_{\kappa}$ . Considérese el modelo abstracto de regresión no paramétrica:

$$X = \Psi(Y) + \epsilon,$$

donde (X, Y) son variables aleatorias definidas sobre un espacio de probabilidad  $(\Omega, \mathcal{F}, P)$  y con valores en  $\mathcal{X} \times \mathcal{Y}, \ \mathcal{X} \subset \mathbb{R}; \ \epsilon$  es una variable aleatoria con valores en los reales y media cero, la cual se supone independiente de  $Y; \ y \ \Psi : \mathcal{Y} \to \mathbb{R}$  es un funcional desconocido. El problema de interés es estimar el funcional de regresión  $\Psi$ sobre la base de datos  $(X_i, Y_i), \ 1 \le i \le n$ , formados por observaciones de (X, Y)independientes e idénticamente distribuidas.

Para esto, sea  $\mathcal{H}_0$  un subespacio lineal de dimensión finita de  $\mathbb{R}^{\mathcal{Y}}$  con base  $G_1, ..., G_m$ , tal que la matriz  $(G_j(Y_i))$ ,  $1 \leq i \leq n, 1 \leq j \leq m$ , tenga rango m. Se denota por  $\mathcal{H} = \mathcal{H}_{\kappa} \oplus \mathcal{H}_0$  el espacio de funciones  $F = F_{\kappa} + F_0$  con  $F_{\kappa} \in \mathcal{H}_{\kappa}$  y  $F_0 \in \mathcal{H}_0$ . De ahora en adelante se supondrá que  $\Psi \in \mathcal{H}$ . Para evaluar cuán cercano está  $F \in \mathcal{H}$  del funcional desconocido  $\Psi$ , supóngase que se ha especificado una función (llamada función de pérdida)  $c : \mathcal{X} \times \mathcal{Y} \times \mathcal{H} \to \mathbb{R}_+$ . El riesgo de aproximar  $\Psi$  por un elemento  $F = F_{\kappa} + F_0 \in \mathcal{H}$  se cuantifica por

$$R(F) = \mathbb{E}\left(c\left(X, Y, F\left(X\right)\right)\right),$$

donde  $\mathbb{E}$  es el valor esperado con respecto a la distribución de (X, Y). El riesgo empírico y el riesgo empírico regularizado se definen, respectivamente como

$$R_{emp}(F) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}\left(c\left(X_{i}, Y_{i}, F\left(X_{i}\right)\right)\right)$$

у

$$R_{\lambda}(F) = R_{emp}(F) + \lambda \|F_{\kappa}\|_{\mathcal{H}_{\kappa}},$$

donde  $\lambda \geq 0$  es una constante dada (hiperparámetro de regularización). La estimación regularizada  $\widehat{\Psi}_{\lambda}$  de  $\Psi$  se define a través del siguiente problema de optimización

$$\widehat{\Psi}_{\lambda} = \arg\min_{F \in \mathcal{H}} R_{\lambda}\left(F\right).$$

Como señalan varios autores, este marco general provee un tratamiento unificado de varios enfoques de regresión no paramétrica, en dependencia de la función de pérdida c especificada (ver, *e.g.*, [25, 34] y referencias citadas en ellos). En particular:

c1) Los métodos de regresión por mínimos cuadrados regularizados corresponden al trabajo con la función de pérdida cuadrática

$$c(X, Y, F(X)) = (X - F(Y))^{2}$$
Específicamente, los métodos de regresión con funciones de bases radiales estándares se obtienen cuando la función de pérdida adoptada es la cuadrática,  $\mathcal{Y}$  es un subespacio de  $\mathbb{R}^d$  (para algún  $d \in \mathbb{N}$ ), y el núcleo tiene forma radial  $\kappa(y, y) = g(||y||)$  para alguna función  $g : \mathbb{R}_+ \to \mathbb{R}$ . Aquí, ||y|| es la norma euclidiana en  $\mathbb{R}^d$ .

c2) Los métodos de regresión de vectores soportes se corresponden con el trabajo con la función de pérdida llamada  $\epsilon$ -insensible

$$c(X, Y, F(X)) = |X - F(Y)|_{\epsilon} = \max\left(|X - F(Y)| - \epsilon, 0\right),$$

donde  $\epsilon \geq 0$  es alguna constante dada. En este caso,  $\mathcal{H}_0$  se toma como el espacio unidimensional generado por la función constante  $G_1(x) \equiv 1$  o por  $\mathcal{H}_0 = \{0\}$ .

# 3.3. Máquinas de vectores soportes funcionales

### 3.3.1. Definición del método FSVR

Como se describió anteriormente, el marco general para la estimación de funciones de regresión a través de la regularización en RKHS puede aplicarse a cualquier especificación del espacio  $\mathcal{Y}$ , del núcleo definido positivo  $\kappa$  en  $\mathcal{Y}$ , de la función de pérdida c y del subespacio de dimensión finita  $\mathcal{H}_0$  de  $\mathbb{R}^{\mathcal{Y}}$ .

Las aplicaciones clásicas de los RKHS para métodos de regresión tratan situaciones en las que  $\mathcal{Y} \subset \mathbb{R}^d$  y, por tanto,  $\mathcal{H} \subset \mathbb{R}^{\mathcal{Y}}$  está constituida por funciones multivariadas  $F : \mathcal{Y} \subset \mathbb{R}^d \to \mathbb{R}$ . Este es el planteamiento para métodos de regresión multivariados, en los cuales el funcional de regresión  $\Psi$  a ser estimado es una *función multivariada*, *i.e.*, una función definida sobre  $\mathbb{R}^d$ .

Por el contrario, los modelos de regresión no paramétricos funcionales tratan casos en que  $\mathcal{Y} \subset \mathbb{R}^{\mathcal{T}}$  es un conjunto de funciones  $y : \mathcal{T} \to \mathbb{R}$ , donde  $\mathcal{T}$  es un conjunto de dimensión infinita. Por ejemplo, en algunas situaciones prácticas de interés,  $\mathcal{T} = [a, b]$ es un intervalo cerrado en  $\mathbb{R}$ , y  $\mathcal{Y}$  es el espacio  $L^2([a, b], \mathbb{R})$  de funciones de cuadrado integrables definidas en [a, b]. Por tanto, en modelos de regresión con datos funcionales el funcional desconocido  $\Psi$  está definido en un espacio normado  $\mathcal{Y}$  de funciones con valores en los reales.

Preda [74] inició la estimación de modelos de regresión funcional en el marco de los RKHS haciendo uso de la función de pérdida cuadrática, y dio por tanto, estimadores de mínimos cuadrados regularizados. Estos son ejemplos de la clase general (c1) descrita en la sección previa. En este capítulo se desarrolla un método para estimar modelos de regresión para datos funcionales adoptando la función de pérdida  $\epsilon$ -insensible, *i.e.* se adapta a datos funcionales el enfoque general (c2) anteriormente mencionado.

Es conocido que bajo condiciones bastante generales de la función de pérdida c(las cuales son satisfechas tanto por la función de pérdida cuadrática como por la  $\epsilon$ insensible), el estimador regularizado  $\widehat{\Psi}_{\lambda}$  en cualquier modelo de regresión abstracto tiene la siguiente forma explícita [34]: para todo  $y \in \mathcal{Y}$ ,

$$\widehat{\Psi}_{\lambda}(y) = \sum_{i=1}^{n} a_{i}\kappa(y_{i}, y) + \sum_{j=1}^{m} b_{j}G_{j}(y)$$

para algunos  $a_i, b_j \in \mathbb{R}$  que dependen sólo del hiperparámetro  $\lambda$  y de las matrices  $K = (\kappa(y_i, y_j)), y G = (G_j(x_i)).$ 

Por tanto, una vez que se tiene el núcleo  $\kappa$ , los cálculos numéricos de los estimadores de regresión en RKHS,  $\widehat{\Psi}_{\lambda}$ , para datos funcionales (*i.e.*, cuando  $\mathcal{Y}$  es un espacio de funciones) son exactamente los mismos que cuando las variables de entrada son multivariadas (*i.e.*, cuando  $\mathcal{Y}$  es un subespacio de algún espacio lineal de dimensión finita  $\mathbb{R}^d$ ).

Además, tanto para la función de pérdida cuadrática como para la  $\epsilon$ -insensible (*i.e.*, para los enfoques de mínimos cuadrados regularizados y de vectores soportes), muchas propiedades de los estimadores  $\widehat{\Psi}_{\lambda}$  en el contexto de datos funcionales pueden ser directamente derivadas de resultados conocidos en el marco de modelos de regresión abstractos. Por ejemplo, si  $\mathcal{X}$  está acotada por una constante  $M_{\mathcal{X}}$  y  $\kappa$  está acotada por una constante  $M_{\kappa}$  entonces, para cualquier  $\lambda \geq 0$ ,

$$R_{emp}\left(\widehat{\Psi}_{\lambda}\left(y\right)\right) - R\left(\widehat{\Psi}_{\lambda}\left(y\right)\right) \to 0$$

exponencialmente en probabilidad cuando  $n \to \infty$  ([73], véase también, [9, 82]). De la misma forma, algunos resultados de convergencia uniforme en probabilidad

$$R_{emp}\left(F\right) - R\left(F\right) \to 0$$

sobre conjuntos de la forma  $\{F : ||F||_{\mathcal{H}_{\kappa}} \leq A\}, A \in \mathbb{R}$ , se discuten en [25].

Sin embargo, la regularización en RKHS tiene especificaciones importantes para el caso de modelos de regresión con datos funcionales. Algunas de ellas, principalmente las relacionadas con la interpretación y construcción de núcleos definidos positivos en espacios funcionales se discuten en el siguiente epígrafe.

### 3.3.2. Núcleos definidos no negativos para espacios funcionales

Los núcleos definidos no negativos en espacios lineales de dimensión finita  $\mathbb{R}^d$ usualmente se construyen sobre la base de operadores diferenciales clásicos o funciones completamente monótonas de orden  $q \in \mathbb{N}$ , de donde se derivan las aproximaciones a funciones multivariadas con splines y funciones de bases radiales [101]. Sin embargo, estos enfoques no son directamente aplicables para obtener núcleos definidos no negativos en espacios de dimensión infinita. Esta dificultad es un reto a los enfoques de regresión para datos funcionales en RKHS.

En esta sección se muestran algunos resultados que permiten la construcción de algunas clases de núcleos definidos no negativos en espacios de funciones. Se supone que  $\mathcal{Y}$  es algún espacio de Hilbert separable (de dimensión infinita). Así, todos los resultados son aplicables a problemas típicos con datos funcionales, en los cuales  $\mathcal{Y} \subset \mathbb{R}^{\mathcal{T}} \ \mathrm{y} \ \mathcal{T}$  es un intervalo en los reales.

El siguiente lema puede demostrarse fácilmente.

Lema 3.1. Sea

$$\varphi\left(y\right) = \mathbb{E}\left(e^{iW(y)}\right)$$

la función característica de un proceso estocástico  $(W(y) : y \in \mathcal{Y})$  de valores reales. Si  $y \to W(y)$  es un funcional lineal, y si para todo  $y \in \mathcal{Y}$ , la distribución de la variable aleatoria W(y) es simétrica alrededor del cero, entonces

$$\kappa\left(y,y\right) = \varphi\left(y-y\right)$$

es una función núcleo, i.e., es definida no negativa sobre  $\mathcal{Y} \times \mathcal{Y}$ .

Demostración:

Sean  $c_j \in \mathbb{R}, y_j \in \mathcal{Y}$ , debe demostrarse que

$$\sum_{r,j} c_r c_j \kappa(y_r, y_j) \ge 0 \quad \forall c_j.$$

Pero se tiene que

$$\begin{split} \sum_{r,j} c_r c_j \kappa(y_r, y_j) &= \sum_{r,j} c_r c_j \mathbb{E} e^{iW(y_r - y_j)} \\ &= \sum_{r,j} c_r c_j \mathbb{E} (e^{iW(y_r)} e^{-iW(y_j)}) \\ &= \mathbb{E} \left| \sum_r c_r e^{iW(y_r)} \right|^2 \ge 0. \end{split}$$

Por otra parte se tiene que

$$\kappa(y_1, y_2) = \mathbb{E}e^{iW(y_1 - y_2)} = \mathbb{E}e^{i(-1)W(y_2 - y_1)} = \kappa(y_2, y_1)$$

por la simetría de la distribución, lo cual finaliza la demostración.  $\square$ 

Una consecuencia directa de este lema es el siguiente hecho bastante conocido: si  $\mathcal{Y}$  es un espacio de Hilbert separable y h > 0, entonces la función Gaussiana

$$\varphi\left(y\right) = e^{-\frac{1}{h}\left\|y\right\|^2}$$

define un núcleo definido no negativo en  $\mathcal{Y}$  (llamado núcleo Gaussiano [82]) a través de

$$\kappa\left(y, y\right) = \varphi\left(y - y\right) = e^{-\frac{1}{h} \|y - y\|^2}.$$

De hecho, es conocido que existe un espacio de probabilidad  $\Omega$  y un proceso estocástico lineal y Gaussiano  $(W(y) : y \in \mathcal{Y})$  en  $\Omega$  con media cero (llamado proceso isonormal  $\mathcal{Y}$ ) tal que

$$\mathbb{E}\left(W\left(y\right)W\left(y\right)\right) = \langle y, y \rangle,$$

donde  $\langle y, y \rangle$  es el producto escalar en  $\mathcal{Y}$ . Por tanto, las condiciones del Lema 3.1 se satisfacen tomando

$$\varphi(y) = \mathbb{E}\left(e^{iW(y)}\right) = e^{-\frac{1}{h}\|y\|^2}.$$

El Lema 3.1 da otra clase interesante de núcleos definidos no negativos, como la que se plantea en el siguiente corolario.

**Corolario 3.1.** Sea  $\mathcal{T}$  un intervalo de  $\mathbb{R}$ ,  $\alpha \in (0,2]$ ,  $h \geq 0$ ,  $y \mathcal{Y}$  un conjunto de funciones medibles  $\mathcal{Y} \subset \mathbb{R}^{\mathcal{T}}$  tal que

$$\left\|y\right\|_{L_{\alpha}} = \left(\int_{\mathcal{T}} |y(t)|^{\alpha} dt\right)^{\frac{1}{\alpha}}$$

es finito para cada  $y \in \mathcal{Y}$ . Entonces,

$$\kappa\left(y,y\right) = e^{-\frac{1}{h}\|y-y\|_{L_{\alpha}}^{\alpha}}$$

es un núcleo definido no negativo en  $\mathcal{Y}$  (el cual se llamará núcleo " $\alpha$ -estable").

De hecho, defínase

$$W(y) = \frac{1}{h} \int_{\mathcal{T}} y(t) \, dZ(t) \,, \qquad (3.2)$$

donde  $(Z(t) : t \in \mathcal{T})$  es un movimiento  $\alpha$ -estable Levy simétrico estándar [81]. Entonces, puede mostrarse que W satisface las condiciones del Lema 3.1.

Nótese que, de manera más general, también como consecuencia del Lema 3.1, si el proceso estocástico  $(Z(t) : t \in \mathcal{T})$  es cualquier semimartingala cuyos incrementos tengan distribuciones simétricas, y si la integral estocástica (3.2) está bien definida para todo  $y \in \mathcal{Y}$ , entonces

$$\kappa(y, y) = \mathbb{E}\left(e^{i(W(y) - W(y))}\right)$$

es un núcleo definido no negativo.

La variedad de núcleos para datos funcionales brindada por estos resultados es un aspecto importante para futuras investigaciones. Como primer paso en esta dirección, en la siguiente sección se explora únicamente el comportamiento de la regresión con vectores soportes para datos funcionales con el núcleo Gaussiano en un conjunto de datos reales.

## 3.4. Resultados en el conjunto de datos "Tecator"

En esta sección se evalúa el comportamiento del método propuesto, FSVR, en el conjunto de datos reales Tecator, descrito en el Capítulo 2. Sobre la base de este conjunto de datos se harán comparaciones con otros métodos de regresión funcionales como son las funciones de bases radiales y el estimador por núcleo de Nadaraya-Watson. Se incluirán además comparaciones con otros métodos de calibración multivariados con que se han trabajado estos datos, como son PLS, SVR y LSSVM.

Los espectros en este conjunto de datos se muestrearon de manera fina, por lo que las curvas obtenidas son curvas muy suaves. Cada función se representa usando una base B-spline de orden 4 y 32 funciones bases. El núcleo adoptado es el Gaussiano. Estudios previos realizados en este conjunto de datos [80] destacan la relevancia que tiene la forma del espectro en la predicción de la propiedad de interés; por esa razón, se usó la norma  $L_2$  de la segunda derivada del espectro dentro del núcleo Gaussiano en FSVR.

La selección de los valores de los hiperparámetros  $(\lambda, \epsilon, h)$  para entrenar el FSVR se realizó explorando una rejilla de búsqueda basada en una validación cruzada con 10 particiones. Para el caso de los estimadores RBF y NWK, se aplicó un criterio de Validación Cruzada Generalizada (GCV, del inglés <u>Generalized Cross-Validation</u>) [101].

La Figura 3.1 muestra las gráficas de los valores predichos de porcentajes de grasa contra los valores reales en el conjunto de prueba para cada uno de los métodos de regresión: FSVR, RBF y NWK. La exactitud de los resultados (capacidad de generalización) es muy buena para todos los métodos, pero FSVR exhibe los mejores.



Figura 3.1: Valores reales contra valores predichos para las estimaciones por a) FSVR, b) RBF, y c) NWK.

La Tabla 3.1 resume los valores de varias medidas de precisión de la predicción en el conjunto de prueba para cada uno de los métodos: RMSE, Error Máximo Absoluto (EMA), Error Relativo Máximo (ERM) y Promedio de Error Relativo (PER). FSVR muestra el mejor comportamiento de predicción entre todas las medidas usadas, mientras que el método NWK es el peor, particularmente con respecto al criterio EMA.

Una representación gráfica del funcional de regresión ajustado se hace conveniente para una mejor comprensión de los diferentes estimadores de regresión en términos de

	RMSE	EMA	ERM	PER
FSVR	0.5451	1.2933	0.1653	0.0276
RBF	0.9279	2.8486	0.3061	0.0512
NWK	1.4563	5.6000	0.2196	0.0642

Tabla 3.1: Exactitud de las predicciones para los diferentes estimadores.

capacidad de predicción y forma de la regresión estimada. Sin embargo, su completa visualización no es posible debido al hecho de que ellos pertenecen a un espacio de dimensión infinita. Una solución a este problema se logra aplicando a la matriz de inter-distancias entre los espectros (distancias calculadas de acuerdo a la norma  $L_2$  de la segunda derivada mencionada anteriormente) una transformación de escalamiento multidimensional al espacio Euclídeo bidimensional. De este modo, cada espectro  $y_i$  se representa (aproximadamente) por un punto  $(u_i, v_i)$  en el plano. Para cada método, la Figura 3.2 muestra la superficie (interpolada) formada por los valores predichos  $\hat{x}_i$  (eje vertical) como función de la representación  $(u_i, v_i)$ , i = 1, ..., n (en el plano horizontal). Los puntos  $(u_i, v_i, x_i)$  que representan los valores de porcentaje de grasa originales  $x_i$  correspondientes al espectro  $y_i$  también se grafican como asteriscos.



Figura 3.2: Estimaciones del funcional de regresión para los estimadores (a) FSVR, (b) RBF, (c) NWK.

En la Figura 3.2 se corrobora también el buen ajuste alcanzado por los tres métodos de estimación considerados, así como la similitud en la forma de sus estimaciones del funcional de regresión.

Sin embargo, deben resaltarse dos ventajas del método FSVR. Primero, como se mencionó anteriormente, muestra mejores resultados de predicción como reporta la Tabla 3.1. Segundo, a diferencia de los métodos funcionales no paramétricos existentes, FSVR alcanza, por su propiedad de raleza, alrededor de un 20% de reducción de la dimensión, y retiene sólo un subconjunto apropiado de vectores soportes (80%) de los datos originales. Los otros dos métodos (RBF y NWK), por el contrario, requieren almacenar todo el conjunto de entrenamiento para predecir muestras futuras.

Se realizó también una comparación del método FSVR con otros métodos de calibración multivariada, no funcionales, usados en este conjunto de datos.

La Figura 3.3 muestra que las FSVR, al igual que los otros métodos no lineales, resuelven el problema de calibración mejor que PLS. Este comportamiento era de esperar producto de la relación no lineal que se conoce que existe entre el contenido de grasa y la información espectral en estos datos. FSVR es más preciso (por un factor de 1.51) que el SVR original no funcional, lo que evidencia que el nuevo método FSVR aprovecha la información funcional proporcionada. La diferencia a favor del FSVR con respecto al LSSVM es sólo de un factor de 1.03 y esto puede atribuirse a que LSSVM es un método simple y requiere la optimización de pocos hiperparámetros en comparación con los otros métodos basados en vectores soportes; por tanto, los modelos pueden ser optimizados de manera más exacta.



Figura 3.3: Valor de RMSEP alcanzado por cada método en el conjunto de prueba.

# Capítulo 4

# REGRESIÓN INVERSA BASADA EN LA DENSIDAD PARA DATOS FUNCIONALES

# 4.1. Introducción

En este capítulo se utilizan los conceptos y las notaciones introducidos en el Capítulo 3 para el problema de la calibración funcional. El propósito es desarrollar un nuevo método de regresión funcional para estimar  $\gamma(Y)$  que también requiere la estimación del modelo de regresión inverso (3.1). Su principal motivación práctica proviene de los problemas de calibración en la quimiometría, específicamente en la espectrometría, donde una variable química X (*e.g.*, concentración) debe predecirse a partir de una función Y (*e.g.*, un espectro). En este contexto, tal modelo inverso representa el proceso físico de generación de los datos en el cual el espectro de salida Y se determina por la concentración química de entrada X, y *e* es una perturbación funcional debida principalmente al proceso de medición. Aunque X y Y puedan tener distribuciones de probabilidad complejas desconocidas, es un supuesto común que la perturbación *e* siga una distribución Gaussiana  $P_0$  y, por tanto, la distribución condicional  $P(\cdot / x)$  de Y dado X = x es una distribución Gaussiana en  $\mathcal{Y}$  con media  $\mu(x)$ . Esto sugiere la siguiente estimación para  $\gamma(y)$ :

$$\widehat{\gamma}(y) = \frac{\sum_{i=1}^{n} \widehat{f}(y \swarrow x_i) x_i}{\widehat{f}_Y(y)},$$

donde  $\hat{f}(y \not< x)$  es una estimación de la densidad  $f(y \not< x)$  de  $P(\cdot \not< x)$  con respecto a la medida  $P_0$  y  $\hat{f}_Y$  es una estimación de la densidad de Y. Este estimador de regresión se denomina Regresión Inversa Basada en la Densidad (DBIR). Si Y fuera una variable con valores en  $\mathbb{R}$ , esto se reduce al enfoque univariado para la calibración propuesto por Lwin [59]. DBIR requiere supuestos más específicos sobre la distribución de la perturbación e en el modelo inverso (3.1) (e.g., distribución Gaussiana) pero tiene un número de características atractivas en contraste con otros enfoques:

- (I) Al igual que otros enfoques no paramétricos, permite aproximar *funciones de regresión no lineales*.
- (II) Al igual que el enfoque FIR, requiere estimar la función media  $\mu$  del modelo inverso y algún operador de covarianza (específicamente,  $\mathbb{V}(e)$ ).  $\mu$  debe ser estimada a través de un método de regresión no paramétrico pero, contrariamente a la regresión funcional no paramétrica, esta función está definida sobre  $\mathbb{R}$  (y no sobre un espacio de Hilbert de dimensión infinita), por lo cual la tarea de estimación es mucho más fácil, así como el problema de ajustar los hiperparámetros (si hay) en el método de regresión no paramétrica adoptado. Además, ningún otro parámetro tiene que ser estimado. En particular, a diferencia de la regresión por núcleo funcional, este enfoque no requiere la selección del parámetro ancho de banda para el núcleo en un espacio de dimensión infinita. Tampoco involucra un número grande de parámetros relacionados con el funcional a optimizar, contrariamente a como sucede, por ejemplo, en el caso del perceptrón funcional. DBIR es por tanto *computacionalmente muy fácil de aplicar*.
- (III) Finalmente, a diferencia del enfoque FIR, se requieren pocos supuestos. En particular,  $\gamma$  no necesita ser una función de un número finito d de proyecciones ni Y tiene que seguir una distribución elíptica (o cualquier otra distribución dada). También nótese que DBIR no requiere el paso adicional del ajuste no paramétrico multivariado (c, paso 3) mencionado en el Capítulo 3.

### 4.2. Regresión inversa basada en la densidad

### 4.2.1. Definición de DBIR en un marco abstracto

Sea (X, Y) un par de variables aleatorias que toma valores en  $\mathbb{R} \times \mathcal{Y}$  donde  $(\mathcal{Y}, \langle ., . \rangle)$ es el espacio de funciones de cuadrados integrables de  $\mathbb{R}$  en  $\mathbb{R}$   $(\mathcal{Y} = L^2([a, b]))$ . Supóngase además que se tienen *n* realizaciones i.i.d. de (X, Y), denotadas por  $(x_i, y_i)_{i=1,...,n}$ . El objetivo es construir, a partir de  $(x_i, y_i)_i$ , un predictor del valor de X correspondiente a un valor futuro observado de Y. Este problema usualmente se trata a través de la estimación de la función de regresión  $\gamma(y) = \mathbb{E}(X \swarrow Y = y)$ .

Para esto, el enfoque funcional de Regresión inversa basado en la densidad ajusta el modelo de regresión inverso

$$Y = F(X) + e, (4.1)$$

donde e es un proceso aleatorio (perturbación o ruido) con media cero, independiente de X, y  $x \to F(x)$  es una función de  $\mathbb{R}$  en  $\mathcal{Y}$ . Como se especificó en la Sección 3.1, este es el modelo de generación de los datos comúnmente supuesto en problemas de calibración [69].

De manera adicional, se adoptan los siguientes supuestos:

1. Existe una medida de probabilidad  $P_0$  en  $\mathcal{Y}$  (no dependiente de X) tal que la medida de probabilidad condicional de Y dado X = x, digamos  $P(\cdot / x)$ , tiene una densidad  $f(\cdot / x)$  con respecto a  $P_0$ ; *i.e.*,

$$P(A \swarrow x) = \int_{A} f(y \measuredangle x) P_{0}(dy)$$

para cualquier conjunto medible A en  $\mathcal{Y}$ ;

2. X es una variable aleatoria continua, *i.e.*, su distribución tiene una densidad  $f_X(x)$  (con respecto a la medida de Lebesgue en  $\mathbb{R}$ ).

Bajo estos supuestos, la función de regresión puede ser escrita como

$$\gamma\left(y\right) = \frac{\int_{\mathbb{R}} f\left(y \nearrow x\right) f_X\left(x\right) x dx}{f_Y\left(y\right)},$$

donde

$$f_{Y}(y) = \int_{\mathbb{R}} f(y \not x) f_{X}(x) dx.$$

Por tanto, dada una estimación  $\widehat{f}(y \neq x)$  de  $f(y \neq x)$ , la siguiente estimación por sustitución de  $\gamma(y)$  puede construirse a partir de las ecuaciones previas:

$$\widehat{\gamma}(y) = \frac{\sum_{i=1}^{n} \widehat{f}(y \neq x_i) x_i}{\widehat{f}_Y(y)}, \qquad (4.2)$$

donde:

$$\widehat{f}_{Y}(y) = \sum_{i=1}^{n} \widehat{f}(y \nearrow x_{i})$$

### 4.2.2. Especificación para el caso Gaussiano

El estimador general dado en la ecuación (4.2) se particulariza en esta sección para el caso en que, para cada  $x \in \mathbb{R}$ ,  $P(\cdot \nearrow x)$  es una medida Gaussiana en  $\mathcal{Y} = \mathcal{L}_2[0, 1]$ .  $P(\cdot \swarrow x)$  está entonces determinada por su correspondiente función de media  $\mu(x) \in \mathcal{Y}$ (la cual es igual a F(x) de acuerdo a la ecuación (4.1)) y un operador de covarianza  $\Gamma$ (no dependiente de x), el cual se supondrá un operador de Hilbert-Schmidt simétrico y positivo en el espacio  $\mathcal{Y}$ . Por tanto, existe una descomposición de valores y vectores propios de  $\Gamma$ ,  $(\varphi_j, \lambda_j)_{j\geq 1}$  tal que  $(\lambda_j)_j$  es una sucesión decreciente de números positivos reales,  $(\varphi_j)_j$  toma valores en  $\mathcal{Y}$  y  $\Gamma = \sum_j \lambda_j \varphi_j \otimes \varphi_j$  donde  $\varphi_j \otimes \varphi_j(h) = \langle \varphi_j, h \rangle \varphi_j$ para cualquier  $h \in \mathcal{Y}$ .

Denótese por  $P_0$  la medida Gaussiana en  $\mathcal{Y}$  con media cero y operador de covarianza  $\Gamma$ . Supóngase que la siguiente condición de regularidad usual se cumple: para cada  $x \in \mathbb{R}$ ,

$$\sum_{j=1}^{\infty} \frac{\mu_j^2\left(x\right)}{\lambda_j} < \infty,$$

donde

$$\mu_{j}(x) = \langle \mu(x), \varphi_{j} \rangle.$$

Entonces,  $P(\cdot \swarrow x)$  y  $P_0$  son medidas Gaussianas equivalentes, y la densidad  $f(\cdot \swarrow x)$  tiene la forma explícita:

$$f(y \swarrow x) = \exp\left\{\sum_{j=1}^{\infty} \frac{\mu_j(x)}{\lambda_j} \left(y_j - \frac{\mu_j(x)}{2}\right)\right\},\,$$

donde  $y_j = \langle y, \varphi_j \rangle$  para todo  $j \ge 1$ . Esto conduce al siguiente algoritmo para estimar  $f(y \swarrow x)$ :

1. Obtención de una estimación  $\hat{\mu}(\cdot)(t)$  de la función  $x \to \mu(x)(t)$  para cada  $t \in \mathbb{R}$ . Esto puede llevarse a cabo a través de cualquier método de regresión no paramétrico estándar basándose en los datos  $(x_i, y_i(t))_{i=1,\dots,n}$ ; *e.g.*, un método de núcleo:

$$\hat{\mu}(x) = \frac{\sum_{i=1}^{n} K\left(\frac{x_i - x}{h}\right) y_i}{\sum_{i=1}^{n} K\left(\frac{x_i - x}{h}\right)}$$
(4.3)

como se propuso en [31] (nótese que, en este caso, el ancho de banda h tiene un valor común para todo t).

2. Obtención de estimaciones  $(\widehat{\varphi}_j, \widehat{\lambda}_j)_j$  de las funciones propias y valores propios  $(\varphi_j, \lambda_j)_j$  de la covarianza  $\Gamma$  sobre la base de la covarianza empírica de los residuos  $\widehat{e}_i = y_i - \widehat{\mu}(x_i), i = 1, ..., n$ 

$$\hat{\Gamma} = \frac{1}{n} \sum_{i=1}^{n} \left( \hat{e}_i - \bar{e} \right) \otimes \left( \hat{e}_i - \bar{e} \right)$$

con  $\bar{e} = \frac{1}{n} \sum_{i=1}^{n} \hat{e}_i$ . Sólo se retienen los primeros p valores propios y funciones propias, donde p = p(n) es un entero dado, menor que n.

3. Estimación de  $f(y \swarrow x)$  por

$$\widehat{f}(y \not x) = \exp\left\{\sum_{j=1}^{p} \frac{\widehat{\mu}_{j}(x)}{\widehat{\lambda}_{j}}\left(\widehat{y}_{j} - \frac{\widehat{\mu}_{j}(x)}{2}\right)\right\},\tag{4.4}$$

donde  $\widehat{\mu}_j(x) = \langle \widehat{\mu}(x), \widehat{\varphi}_j \rangle \ \text{y} \ \widehat{y}_j = \langle y, \widehat{\varphi}_j \rangle.$ 

Finalmente, con la sustitución de (4.4) en (4.2) se obtiene una estimación  $\widehat{\gamma}(y)$  de  $\gamma(y)$ , que es el estimador DBIR.

# 4.3. Consistencia del estimador

En esta sección se demuestran algunas propiedades asintóticas de los estimadores propuestos en los pasos (1)-(3), así como los resultados de consistencia para el estimador DBIR,  $\hat{\gamma}(y)$ . La demostración de la consistencia del estimador que se muestra sigue esencialmente los mismos pasos requeridos para el cálculo del estimador DBIR.

Antes de mostrar los resultados, se hace necesario recordar y definir algunas notaciones útiles.

### 4.3.1. Notaciones

Denótese por:

- $\mathcal{H}$  un espacio de Hilbert;
- *m* la función tal que  $\mu(x)(\cdot) = \frac{m(x)}{f_X(x)}$ ;

- $r \in \mathcal{H} : x \to \mu(x)(\cdot), \quad r_{\eta_n}(x) = \frac{m(x)}{f_{X,\eta_n}(x)}, \quad f_{X,\eta_n}(x) = \max(f_X(x),\eta_n);$
- $\eta_n > 0$  sucesión numérica que converge a cero;
- $\hat{\mu}(x)(\cdot)$  estimación de  $\mu(x)(\cdot)$  usando el estimador por núcleo de Nadaraya-Watson:

$$\hat{\mu}(x)(\cdot) = \frac{\sum_{i=1}^{n} K\left(\frac{x_i - x}{h}\right) y_i(\cdot)}{\sum_{i=1}^{n} K\left(\frac{x_i - x}{h}\right)}$$

- $\hat{r}_{\eta_n} \in \mathcal{H} : x \to \hat{\mu}(x)(\cdot), \quad \hat{r}_{\eta_n}(x) = \frac{\hat{m}(x)}{\hat{f}_{X,\eta_n}(x)}, \quad \hat{f}_{X,\eta_n}(x) = \max(\hat{f}_X(x),\eta_n);$
- $\hat{m}(x) = \sum_{i=1}^{n} y_i K\left(\frac{x_i x}{h}\right);$
- $\hat{f}_X(x) = \sum_{i=1}^n K\left(\frac{x_i x}{h}\right);$
- $\Gamma = \mathbb{E} \left( e \otimes (e \mathbb{E}(e)) \right)$  el operador de covarianza de los errores e = y r(x), que se supone un operador de Hilbert-Schmidt. Por tanto,  $\Gamma$  admite la descomposición espectral  $(\lambda_j, \varphi_j)_j$  donde  $(\lambda_j)_j$  es una sucesión decreciente de valores positivos que tienden a cero y  $(\varphi_j)_j$  es una base ortonormal de  $\mathcal{Y}$ ;
- $\Gamma_n = \frac{1}{n} \sum_{i=1}^n e_i \otimes e_i$ , con  $e_i = y_i r(x_i)$ , el operador de covarianza empírica de *e*.

### 4.3.2. Parte 1: Convergencia de $\hat{r}_{\eta_n}(x)$ y razón de convergencia

El primer paso en el método DBIR es la estimación de la media condicional  $\mu(x)$  a través del estimador de Nadaraya-Watson  $\hat{\mu}(x)$ , definido como en (4.3). Los resultados de convergencia, así como la razón de convergencia se obtienen bajo los siguientes supuestos:

- (A1)  $f_X$  y r son  $\mathcal{C}^k$ ;
- (A2) K es un núcleo de orden k con soporte compacto,  $k \ge 2$ ;
- (A3) existen  $d_1$  y  $d_2$  tales que  $\sup_x \left| f_X^{(k)}(x) \right| < d_1$  y  $\sup_x \left\| r^{(k)}(x) \right\| < d_2$ ;

(A4) 
$$h = O(n^{-c_1}), \eta_n = O(n^{-c_2}), c_1, c_2 > 0, \frac{c_2}{k} + \frac{1}{4k} < c_1 < \frac{1}{4} - c_2;$$

- (A5) existe  $b \in \mathbb{R}^+$  tal que  $\sup_x ||r(x)|| \le b$ .
- (A6)  $n\mathbb{E}\sup_{x} \|r(x)\|^2 I_{\{f_X(x) < \eta_n\}} \xrightarrow{n \to +\infty, P} 0.$

**Proposición 4.1.** Bajo los supuestos (A1)-(A5), se tiene que:

$$\sup_{x} \|r(x) - \hat{r}_{\eta_n}(x)\| = O_P\left(n^{c_2 - c_1k} + \left(\frac{\log n}{n^{1 - 2(c_1 + c_2)}}\right)^{1/2}\right).$$

Demostraci'on:

Para la demostración de esta proposición se hace necesario el uso de los siguientes Lemas. Sus demostraciones aparecen en los artículos citados.

Lema 4.1 ([77]). Bajo los supuestos (A1)-(A3), se tiene que:

$$\sup_{x} \left| \hat{f}_X(x) - f_X(x) \right| = O_P\left( h^k + \frac{\sqrt{\log n}}{\sqrt{nh}} \right).$$

Lema 4.2 ([110]). Bajo los supuestos (A1)-(A3), se tiene que:

$$\sup_{x} \|m(x) - \hat{m}(x)\| = O_P\left(h^k + \frac{\sqrt{\log n}}{\sqrt{nh}}\right)$$

La demostración de la Proposición 4.1 se muestra a continuación: Para todo y, se tiene que

$$\begin{aligned} \|r(x) - \hat{r}_{\eta_n}(x)\| &= \|r(x) - r_{\eta_n}(x) + r_{\eta_n}(x) - \hat{r}_{\eta_n}(x)\| \\ &\leq \|r(x) - r_{\eta_n}(x)\| + \|r_{\eta_n}(x) - \hat{r}_{\eta_n}(x)\| \end{aligned}$$

por lo tanto

$$\sup_{x} \|r(x) - \hat{r}_{\eta_n}(x)\| \le \sup_{x} \|r(x) - r_{\eta_n}(x)\| + \sup_{x} \|r_{\eta_n}(x) - \hat{r}_{\eta_n}(x)\|.$$
(4.5)

Además se cumple que

$$||r_{\eta_n}(x)|| \leq ||r(x)||,$$
 (4.6)

$$\|r(x) - r_{\eta_n}(x)\|^2 = \|r(x) - r_{\eta_n}(x)\|^2 I_{\{f_X(x) < \eta_n\}} \le 4 \|r(x)\|^2 I_{\{f_X(x) < \eta_n\}}, \quad (4.7)$$

con lo cual, usando la desigualdad de Markov, se obtiene

$$P\left(n\sup_{x} \|r(x) - r_{\eta_n}(x)\| > c\right) \le \frac{4}{c} n \mathbb{E}\left[\sup_{x} \|r(x)\|^2 I_{\{f_X(x) < \eta_n\}}\right],$$

que, por el supuesto (A6), da como resultado

$$\sup_{x} \|r(x) - r_{\eta_n}(x)\| = O_P(1/n).$$
(4.8)

Por otra parte, se tiene que

$$\begin{aligned} \|r_{\eta_n}(x) - \hat{r}_{\eta_n}(x)\| &= \left\| \frac{r_{\eta_n}(x)}{\hat{f}_{X,\eta_n}(x)} \left( \hat{f}_{X,\eta_n}(x) - f_{X,\eta_n}(x) \right) + \frac{1}{f_{X,\eta_n}(x)} \left( m(x) - \hat{m}(x) \right) \right\| \\ &\leq \left\| \frac{\|r(x)\|}{\eta_n} \left| \hat{f}_{X,\eta_n}(x) - f_{X,\eta_n}(x) \right| + \frac{1}{\eta_n} \left\| m(x) - \hat{m}(x) \right\|, \end{aligned}$$
(4.9)

donde

$$\left|\hat{f}_{X,\eta_n}(x) - f_{X,\eta_n}(x)\right| \le \left|\hat{f}_X(x) - f_X(x)\right| \le \sup_x \left|\hat{f}_X(x) - f_X(x)\right|$$
 (4.10)

entonces, tomando el supremo en (4.9) y, usando el supuesto (A5) queda

$$\sup_{x} \|r_{\eta_n}(x) - \hat{r}_{\eta_n}(x)\| \le \frac{b}{\eta_n} \sup_{x} \left| \hat{f}_X(x) - f_X(x) \right| + \frac{1}{\eta_n} \sup_{x} \|m(x) - \hat{m}(x)\|.$$

Por el Lema 4.1, el Lema 4.2 y el supuesto (A4), se obtiene que

$$\sup_{x} \|r_{\eta_n}(x) - \hat{r}_{\eta_n}(x)\| = O_P\left(n^{c_2 - c_1k} + \left(\frac{\log n}{n^{1 - 2(c_1 + c_2)}}\right)^{1/2}\right)$$
(4.11)

De la ecuación (4.5) y los resultados previos dados en las ecuaciones (4.8) y (4.11), puede concluirse que

$$\sup_{x} \|r(x) - \hat{r}_{\eta_n}(x)\| = O_P\left(n^{c_2 - c_1k} + \left(\frac{\log n}{n^{1 - 2(c_1 + c_2)}}\right)^{1/2}\right).\square$$

# 4.3.3. Parte 2: Convergencia de $\hat{\Gamma}$

El segundo paso del método DBIR consiste en la estimación del operador de covarianza del error  $\Gamma$  sobre la base de las estimaciones de los residuos. La  $\sqrt{n}$ -consistencia

de este estimador garantiza la consistencia de su correspondiente descomposición espectral (*i.e.*, autovalores y autofunciones) usando los resultados que se muestran en [8]. Para obtener la consistencia del estimador del operador de covarianza el siguiente supuesto, además de los supuestos (A1)-(A6), es necesario:

(A7)  $\mathbb{E}(||e||^4) < +\infty;$ 

**Proposición 4.2.** Bajo los supuestos (A1)-(A4), (A6)-(A7) se tiene:

$$\left\|\Gamma - \hat{\Gamma}\right\| = O_P\left(1/\sqrt{n}\right).$$

Demostración:

Para la demostración de esta proposición se requiere el uso adicional del siguiente Lema cuya demostración puede encontrarse en el artículo citado.

Lema 4.3 ([14]). Sea Z una variable aleatoria en un espacio de Hilbert con operador de covarianza  $\Gamma_Z \ y \mathbb{E}(||Z||^4) < +\infty$  entonces,

$$\mathbb{E}\left(\left\|\Gamma_{Z} - \Gamma_{Z}^{n}\right\|^{2}\right) \leq \frac{\mathbb{E}\left(\left\|Z\right\|^{4}\right)}{n}$$

donde  $\Gamma_Z^n = \frac{1}{n} \sum_{i=1}^n (Z_i - \overline{Z}) \otimes (Z_i - \overline{Z}), y \text{ los } Z_i \text{ son independientes e idénticamente distribuidas (i.i.d) según la distribución de Z.$ 

Por la definición del estimador  $\hat{\Gamma}$  se tiene que:

$$\hat{\Gamma} = \frac{1}{n} \sum_{i=1}^{n} \hat{e}_i \otimes \hat{e}_i = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{r}_{\eta_n}(x_i)) \otimes (y_i - \hat{r}_{\eta_n}(x_i))$$

$$= \frac{1}{n} \sum_{i=1}^{n} (y_i - r(x_i) + r(x_i) - r_{\eta_n}(x_i) + r_{\eta_n}(x_i) - \hat{r}_{\eta_n}(x_i)) \otimes (y_i - r(x_i) + r(x_i) - r_{\eta_n}(x_i) + r_{\eta_n}(x_i) - \hat{r}_{\eta_n}(x_i))$$

el cual puede expresarse como

$$\hat{\Gamma} = \Gamma_n + T_1 + T_1^* + T_2 + T_2^* + T_3 + T_4 + T_4^* + T_5$$

donde

$$T_{1} = \frac{1}{n} \sum_{i=1}^{n} (y_{i} - r(x_{i})) \otimes (r(x_{i}) - r_{\eta_{n}}(x_{i})),$$

$$T_{2} = \frac{1}{n} \sum_{i=1}^{n} (y_{i} - r(x_{i})) \otimes (r_{\eta_{n}}(x_{i}) - \hat{r}_{\eta_{n}}(x_{i})),$$

$$T_{3} = \frac{1}{n} \sum_{i=1}^{n} (r(x_{i}) - r_{\eta_{n}}(x_{i})) \otimes (r(x_{i}) - r_{\eta_{n}}(x_{i})),$$

$$T_{4} = \frac{1}{n} \sum_{i=1}^{n} (r(x_{i}) - r_{\eta_{n}}(x_{i})) \otimes (r_{\eta_{n}}(x_{i}) - \hat{r}_{\eta_{n}}(x_{i})),$$

$$T_{5} = \frac{1}{n} \sum_{i=1}^{n} (r_{\eta_{n}}(x_{i}) - \hat{r}_{\eta_{n}}(x_{i})) \otimes (r_{\eta_{n}}(x_{i}) - \hat{r}_{\eta_{n}}(x_{i})),$$

y  $T_1^*, T_2^*, T_4^*$  son los operadores auto-adjuntos de  $T_1, T_2, T_4$  respectivamente.

Además se sabe que

$$\left\| \Gamma - \hat{\Gamma} \right\| \le \left\| \Gamma - \Gamma_n \right\| + 2 \left\| T_1 \right\| + 2 \left\| T_2 \right\| + \left\| T_3 \right\| + 2 \left\| T_4 \right\| + \left\| T_5 \right\|,$$
(4.12)

por lo que en lo adelante se procederá a acotar cada uno de los términos del miembro derecho de la desigualdad.

Puede deducirse fácilmente del Lema 4.3 que

$$\|\Gamma - \Gamma_n\| = O_P\left(1/\sqrt{n}\right). \tag{4.13}$$

De las ecuaciones (4.6), (4.9) y (4.10) se obtiene

$$\sqrt{n} \|T_5\| \leq \sqrt{n} \frac{1}{n} \sum_{i=1}^n \|r_{\eta_n}(x_i) - \hat{r}_{\eta_n}(x_i)\|^2 \leq \sqrt{n} \sup_x \|r(x) - \hat{r}_{\eta_n}(x)\|^2$$

lo cual, por la Proposición 4.1, da como resultado

$$\sqrt{n} \|T_5\| = O_P \left( n^{2c_2 + 1/2} \left[ n^{-c_1k} + \frac{\sqrt{\log n}}{n^{(-c_1 + 1/2)}} \right]^2 \right)$$

y consecuentemente,

$$||T_5|| = o_P\left(1/\sqrt{n}\right). \tag{4.14}$$

$$\begin{split} \sqrt{n} \|T_2\| &\leq \sqrt{n} \frac{1}{n} \sum_{i=1}^n \|e_i\| \|r_{\eta_n}(x_i) - \hat{r}_{\eta_n}(x_i)\| \\ &\leq \frac{\sqrt{n}}{\eta_n} \sup_x \left| \hat{f}_X(x) - f_X(x) \right| \left( \frac{1}{n} \sum_{i=1}^n \|e_i\| \|r(x_i)\| \right) \\ &\quad + \frac{\sqrt{n}}{\eta_n} \sup_x \|m(x) - \hat{m}(x)\| \left( \frac{1}{n} \sum_{i=1}^n \|e_i\| \right) \\ &\leq \frac{\sqrt{n}}{\eta_n} \sup_x \left| \hat{f}_X(x) - f_X(x) \right| \times b \left( \frac{1}{n} \sum_{i=1}^n \|e_i\| \right) \\ &\quad + \frac{\sqrt{n}}{\eta_n} \sup_x \|m(x) - \hat{m}(x)\| \left( \frac{1}{n} \sum_{i=1}^n \|e_i\| \right). \end{split}$$

De aquí se tiene que, para cualquier c > 0 (usando la desigualdad de Markov, el supuesto (A7) y la desigualdad de Cauchy-Schwarz), se cumple

$$P\left(\sqrt{n} \|T_2\| > c\right) \leq P\left(\sup_x |f_X(x) - f_X(x)| \left(\frac{1}{n} \sum_{i=1}^n \|e_i\|\right) > \frac{c\eta_n}{2b\sqrt{n}}\right)$$
$$+ P\left(\sup_x \|m(x) - \hat{m}(x)\| \left(\frac{1}{n} \sum_{i=1}^n \|e_i\|\right) > \frac{c\eta_n}{2\sqrt{n}}\right)$$
$$\leq \frac{2\sqrt{n}}{\eta_n c} \left[b\mathbb{E}\left(\|e\|\sup_x \left|\hat{f}_X(x) - f_X(x)\right|\right)$$
$$+ \mathbb{E}\left(\|e\|\sup_x \|m(x) - \hat{m}(x)\|\right)\right]$$
$$\leq \frac{2\sqrt{n}\sqrt{\mathbb{E}(\|e\|^2)}}{\eta_n c} \left[b\sqrt{\mathbb{E}\left(\sup_x \left|\hat{f}_X(x) - f_X(x)\right|\right)^2}\right]$$
$$+ \sqrt{\mathbb{E}\left(\sup_x \|m(x) - \hat{m}(x)\|\right)^2}\right]$$

pero por el supuesto (A4), los Lemas 4.1 y 4.2,  $\frac{\sqrt{n}}{\eta_n} \sup_x \left| \hat{f}_X(x) - f_X(x) \right| = o_P(1)$  y  $\frac{\sqrt{n}}{\eta_n} \sup_x ||m(x) - \hat{m}(x)|| = o_P(1)$ ; por tanto

$$\frac{\sqrt{n}}{\eta_n} \sqrt{\mathbb{E}\left(\sup_{y} \left| \hat{f}_X(x) - f_X(x) \right| \right)^2} \xrightarrow[n \to +\infty]{} 0$$
$$\frac{\sqrt{n}}{\eta_n} \sqrt{\mathbb{E}\left(\sup_{x} \left\| m(x) - \hat{m}(x) \right\| \right)^2} \xrightarrow[n \to +\infty]{} 0$$

lo que implica

$$||T_2|| = o_P\left(1/\sqrt{n}\right). \tag{4.15}$$

De la definición de  $T_3$  se tiene que

$$n ||T_3|| \le \sum_{i=1}^n ||r(x_i) - r_{\eta_n}(x_i)||^2.$$

Aplicando la desigualdad de Markov

$$P\left(\sum_{i=1}^{n} \|r(x_{i}) - r_{\eta_{n}}(x_{i})\|^{2} > c\right) \leq \frac{n}{c} \mathbb{E}\left(\|r(x) - r_{\eta_{n}}(x)\|^{2}\right), \quad \forall c \in \mathbb{R},$$

lo cual por la ecuación (4.7) da como resultado

$$P\left(\sum_{i=1}^{n} \|r(x_i) - r_{\eta_n}(x_i)\|^2 > c\right) \le \frac{4}{c} n \mathbb{E}\left(\|r(x)\|^2 I_{\{f_X(x) < \eta_n\}}\right),$$

por tanto, por el supuesto (A6), se deduce que

$$||T_3|| = o_P(1/n). (4.16)$$

De la definición de  $T_4$ , la ecuación (4.9) y el supuesto (A5), se tiene que

$$\begin{split} \sqrt{n} \|T_4\| &\leq \sqrt{n} \frac{1}{\eta_n} \sup_x \left| \hat{f}_X(x) - f_X(x) \right| \left( \frac{1}{n} \sum_{i=1}^n \|r(x_i)\| \|r(x_i) - r_{\eta_n}(x_i)\| \right) \\ &+ \sqrt{n} \frac{1}{\eta_n} \sup_x \|m(x) - \hat{m}(x)\| \left( \frac{1}{n} \sum_{i=1}^n \|r(x_i) - r_{\eta_n}(x_i)\| \right) \\ &\leq \frac{\sqrt{n}}{\eta_n} \sup_x \left| \hat{f}_X(x) - f_X(x) \right| \times b \sup_x \|r(x) - r_{\eta_n}(x)\| \\ &+ \frac{\sqrt{n}}{\eta_n} \sup_x \|m(x) - \hat{m}(x)\| \times \sup_x \|r(x) - r_{\eta_n}(x)\|. \end{split}$$

у

Entonces, usando los Lemas 4.1 y 4.2, el supuesto (A4) y la Proposición 4.1, se obtiene que

$$\frac{\sqrt{n}}{\eta_n} \sup_x \left| \hat{f}_X(x) - f_X(x) \right|, \quad \frac{\sqrt{n}}{\eta_n} \sup_x \left\| m(x) - \hat{m}(x) \right\|, \quad \sup_x \left\| r(x) - r_{\eta_n}(x) \right\|$$

son todos  $o_P(1)$ , lo que implica que

$$||T_4|| = o_P\left(1/\sqrt{n}\right). \tag{4.17}$$

De la definición de  $T_1$ ,

$$\sqrt{n} \|T_1\| \le \frac{\sqrt{n}}{n} \sum_{i=1}^n \|e_i\| \|r(x_i) - r_{\eta_n}(x_i)\|.$$

Por el supuesto (A7),

$$P\left(\frac{\sqrt{n}}{n}\sum_{i=1}^{n}\|e_i\|\|r(x_i) - r_{\eta_n}(x_i)\| > c\right) \leq \frac{\sqrt{n}}{c}\mathbb{E}\left(\|e\|\|r(x) - r_{\eta_n}(x)\|\right), \quad \forall c \in \mathbb{R}$$
$$\leq \frac{1}{c}\sqrt{\mathbb{E}\|e\|^2}\sqrt{n\mathbb{E}\|r(x) - r_{\eta_n}(x)\|^2}$$
$$\leq \frac{2}{c}\sqrt{\mathbb{E}\|e\|^2}\sqrt{n\mathbb{E}\|r(x)\|^2}I_{f_X(x)>\eta_n}$$

lo cual, por el supuesto (A6), tiende a cero. Por tanto,

$$||T_1|| = o_P\left(1/\sqrt{n}\right). \tag{4.18}$$

Con los resultados alcanzados en la ecuaciones (4.13), (4.18), (4.15), (4.16), (4.17) y (4.14), junto con la ecuación (4.12) puede concluirse que

$$\left\| \Gamma - \hat{\Gamma} \right\| = O_P\left( 1/\sqrt{n} \right) . \Box$$

# 4.3.4. Parte 3: Convergencia de $\hat{f}(y \neq x)$

El último paso del método DBIR es estimar la densidad condicional  $f(y \not/ x)$  a partir de la descomposición espectral de  $\hat{\Gamma}$ . Como se explicó anteriormente, este resultado se deriva a partir de un Teorema demostrado en [8], por tanto, se requerirán algunas notaciones y supuestos técnicos hechos en ese trabajo. A continuación se enuncia, a modo de Lema, el resultado que se necesita usar: Lema 4.4 ([110]). Sea  $\Gamma$  y  $\tilde{\Gamma}$  dos operadores lineales compactos y auto-adjuntos definidos en un espacio de Hilbert, y  $(\lambda_j, \varphi_j)_{j \in \mathbb{N}}$ ,  $(\tilde{\lambda}_j, \tilde{\varphi}_j)_{j \in \mathbb{N}}$  las respectivas sucesiones decrecientes de autovalores y las sucesiones de autovectores ortonormales. Entonces, para todo  $j \in \mathbb{N}$ 

 $i) |\lambda_{j} - \tilde{\lambda}_{j}| \leq ||\Gamma - \tilde{\Gamma}||;$   $ii) ||\varphi_{j} - \tilde{\varphi}_{j}|| \leq a_{j} ||\Gamma - \tilde{\Gamma}|| donde$   $a_{j} = \begin{cases} \frac{2\sqrt{2}}{\lambda_{1} - \lambda_{2}} & si \ j = 1\\ \frac{2\sqrt{2}}{\min(\lambda_{j-1} - \lambda_{j}, \lambda_{j} - \lambda_{j+1})} & si \ j \geq 2 \end{cases}.$ 

Supóngase ahora que los siguientes supuestos se satisfacen:

(A8) todos los  $(\lambda_j)_j$  son distintos;

(A9) 
$$\lim_{n \to +\infty} p = +\infty$$
;

- (A10) lím<sub> $n\to+\infty$ </sub>  $\frac{\sum_{j=1}^{p} a_j}{\lambda_p \sqrt{n}} = 0$  donde  $(a_j)_j$  se definen como en el Lema 4.4;
- (A11)  $\lim_{n \to +\infty} \frac{p}{\lambda_p} \times \left( n^{c_2 c_1 k} + \left( \frac{\log n}{n^{1 2(c_1 + c_2)}} \right)^{1/2} \right) = 0 \text{ y } \lim_{n \to +\infty} \frac{p}{\lambda_p^2 \sqrt{n}} = 0 \text{ ;}$

(A12) para todo  $y \in \mathcal{Y}, \sum_{j\geq 1} \sup_{x} \left| \frac{\mu_j(x)}{\lambda_j} \left( y_j - \frac{\mu_j(x)}{2} \right) \right| < +\infty.$ 

**Proposición 4.3.** Bajo los supuestos (A1)-(A12), para cualquier  $y \in \mathcal{Y}$ , se tiene:

$$\sup_{x} \left| f(y \not x) - \hat{f}(y \not x) \right| = o_P(1) \,.$$

Demostración:

Denote por  $\mathcal{E}(x)$  la diferencia

$$\left|\sum_{j=1}^{p} \frac{\hat{\mu}_j(x)}{\hat{\lambda}_j} \left( \hat{y}_j - \frac{\hat{\mu}_j(x)}{2} \right) - \sum_{j \ge 1} \frac{\mu_j(x)}{\lambda_j} \left( y_j - \frac{\mu_j(x)}{2} \right) \right|.$$

Entonces,  $\mathcal{E}(x) \leq \mathcal{E}_1(x) + \mathcal{E}_2(x)$ , donde

$$\mathcal{E}_1(x) = \left| \sum_{j=p+1}^{+\infty} \frac{\mu_j(x)}{\lambda_j} \left( y_j - \frac{\mu_j(x)}{2} \right) \right|$$

у

$$\mathcal{E}_2(x) = \left| \sum_{j=1}^p \left[ \frac{\hat{\mu}_j(x)}{\hat{\lambda}_j} \left( \hat{y}_j - \frac{\hat{\mu}_j(x)}{2} \right) - \frac{\mu_j(x)}{\lambda_j} \left( y_j - \frac{\mu_j(x)}{2} \right) \right] \right|$$

Por el supuesto (A9),  $\lim_{n \to +\infty} p = +\infty$  y como  $\sum_{j \ge 1} \sup_x \left| \frac{\mu_j(x)}{\lambda_j} \left( y_j - \frac{\mu_j(x)}{2} \right) \right|$  es finito (por el supuesto (A12)),  $\lim_{n \to +\infty} \sup_x \mathcal{E}_1(x) = 0.$ 

 $\mathcal{E}_2$  puede dividirse en 4 partes:  $\mathcal{E}_2(x) \leq A(x) + B(x) + C(x) + D(x)$  donde

• 
$$A(x) = \left| \sum_{j=1}^{p} \frac{\mu_j(x)}{\lambda_j} (y_j - \hat{y}_j) \right|;$$

• 
$$B(x) = \left| \sum_{j=1}^{p} \frac{\mu_j(x)}{2\lambda_j} (\mu_j(x) - \hat{\mu}_j(x)) \right|;$$

• 
$$C(x) = \left| \sum_{j=1}^{p} \frac{\hat{y}_j - \hat{\mu}_j(x)/2}{\lambda_j} (\mu_j(x) - \hat{\mu}_j(x)) \right|;$$

•  $D(x) = \left| \sum_{j=1}^{p} \left( \frac{1}{\lambda_j} - \frac{1}{\hat{\lambda}_j} \right) \hat{\mu}_j(x) \left( \hat{y}_j - \frac{\hat{\mu}_j(x)}{2} \right) \right|.$ 

### Consistencia de A

Como  $\sum_{j} \mu_{j}(x)^{2} = ||r(x)||^{2}$ , entonces, para todo  $j, |\mu_{j}(x)| \leq ||r(x)||$ . Esta desigualdad, así como el hecho de que  $\lambda_{j} \geq \lambda_{p}$  para  $j = 1, \ldots, p$  y  $|y_{j} - \hat{y}_{j}| = |\langle y, \varphi_{j} - \hat{\varphi}_{j} \rangle_{\mathcal{Y}}| \leq ||y|| ||\varphi_{j} - \hat{\varphi}_{j}||$ , conduce a que, para todo x,

$$A(x) \le \frac{\|r(x)\|}{\lambda_p} \|y\| \sum_{j=1}^p \|\varphi_j - \hat{\varphi}_j\|.$$

Entonces, usando el Lema 4.4 (con los operadores  $\Gamma$  y  $\hat{\Gamma}$ ) y los supuestos (A5),(A8), se tiene que:

$$\sup_{x} A(x) \le b \|y\| \|\Gamma - \hat{\Gamma}\| \frac{\sum_{j=1}^{p} a_j}{\lambda_p}$$

lo cual, por la Proposición 4.2 y el supuesto (A10) implica que  $\sup_x A(x) = o_P(1)$ .

### Consistencia de B

Por los mismos argumentos que los usados para A, se tiene que

$$B(x) \le \frac{\|r(x)\|}{2\lambda_p} \sum_{j=1}^p |\mu_j(x) - \hat{\mu}_j(x)|$$

Por otra parte, para cualquier j y cualquier x, se cumple

$$|\mu_j(x) - \hat{\mu}_j(x)| \le ||r(x)|| ||\varphi_j - \hat{\varphi}_j|| + ||\hat{\varphi}_j|| ||r(x) - \hat{r}_{\eta_n}(x)||$$

y por tanto, para cada j,

$$\sup_{x} |\mu_{j}(x) - \hat{\mu}_{j}(x)| \le b \, a_{j} \|\Gamma - \hat{\Gamma}\| + \sup_{x} \|r(x) - \hat{r}_{\eta_{n}}(x)\|$$

donde la primera parte se obtiene por el Lema 4.4 y el supuesto (A5). Por tanto,

$$\sup_{x} B(x) \le \frac{b^2}{2} \|\Gamma - \hat{\Gamma}\| \frac{\sum_{j=1}^p a_j}{\lambda_p} + \frac{b}{2} \frac{p}{\lambda_p} \sup_{x} \|r(x) - \hat{r}_{\eta_n}(x)\|$$

donde el primer término del miembro derecho de la desigualdad es  $o_P(1)$  (Proposición 4.2 y supuesto (A10)) y el segundo término es

$$O_P\left(\frac{p}{\lambda_p} \times \left(n^{c_2-c_1k} + \left(\frac{\log n}{n^{1-2(c_1+c_2)}}\right)^{1/2}\right)\right)$$

por la Proposición 4.1, resultando la suma de ellos  $o_P(1)$  (supuesto (A11)). Entonces,

$$\sup_{x} B(x) = o_P(1) \,.$$

### Consistencia de C

De  $|\hat{y}_j| \le ||y|| \le ||\hat{\mu}_j(x)| \le ||\hat{r}_{\eta_n}(x)||$ , se tiene que

$$C(x) \le \frac{\|y\| + \|\hat{r}_{\eta_n}(x)\|}{\lambda_p} \sum_{j=1}^p |\mu_j(x) - \hat{\mu}_j(x)|.$$

De forma análoga a como se hizo para B, se demuestra que  $\sup_x \frac{\sum_{j=1}^p |\mu_j(x) - \hat{\mu}_j(x)|}{\lambda_p} = o_P(1)$ . Además, para cualquier x,  $\|\hat{r}_{\eta_n}(x)\| = \|r(x)\| + o_P(1)$  por la Proposición 4.1 y por tanto,  $\sup_x \|\hat{r}_{\eta_n}(x)\| \le b + o_P(1)$ . Todo esto conduce a  $\sup_x C(x) = o_P(1)$ .

#### Consistencia de D

Al usar los mismos argumentos que para C, se tiene que

$$D(x) \leq \|\hat{r}_{\eta_n}(x)\| \left( \|y\| + \frac{\|\hat{r}_{\eta_n}(x)\|}{2} \right) \sum_{j=1}^{p} \left| \frac{1}{\lambda_j} - \frac{1}{\hat{\lambda}_j} \right|$$
  
$$\leq \left( \|\hat{r}_{\eta_n}(x)\| \|y\| + \frac{\|\hat{r}_{\eta_n}(x)\|^2}{2} \right) \frac{p\|\Gamma - \hat{\Gamma}\|}{\lambda_p \hat{\lambda}_p}.$$

Ahora, partiendo de que  $\hat{\lambda}_p \ge \left|\lambda_p - |\lambda_p - \hat{\lambda}_p|\right|$  se obtiene

$$\frac{p\|\Gamma - \hat{\Gamma}\|}{\lambda_p \hat{\lambda}_p} \le \frac{p\|\Gamma - \hat{\Gamma}\|}{\left|\lambda_p^2 - \lambda_p |\lambda_p - \hat{\lambda}_p|\right|} = \frac{\frac{p}{\lambda_p^2} \|\Gamma - \hat{\Gamma}\|}{\left|1 - \frac{|\lambda_p - \hat{\lambda}_p|}{\lambda_p}\right|},$$

lo cual, usando la expansión  $\frac{1}{1-y} = \sum_{j=0}^\infty y^j, \forall |y| < 1$ y el Lema 4.4 da como resultado

$$\frac{\frac{p}{\lambda_p^2} \|\Gamma - \hat{\Gamma}\|}{\left|1 - \frac{|\lambda_p - \hat{\lambda}_p|}{\lambda_p}\right|} = \frac{p}{\lambda_p^2} \|\Gamma - \hat{\Gamma}\| \left(1 + \frac{|\lambda_p - \hat{\lambda}_p|}{\lambda_p} + o\left(\frac{|\lambda_p - \hat{\lambda}_p|}{\lambda_p}\right)\right)$$
$$= \frac{p}{\lambda_p^2} O_P\left(\frac{1}{\sqrt{n}}\right) + \frac{p}{\lambda_p^3} O_P\left(\frac{1}{n}\right) + \frac{p}{\lambda_p^2} O_P\left(\frac{1}{\sqrt{n}}\right) o_P\left(\frac{1}{\sqrt{n}\lambda_p}\right).$$

Por el supuesto (A11) se tiene que el primer término del miembro derecho es  $o_P(1)$ y el segundo término es equivalente a  $O_P\left(\frac{p}{\lambda_p^2\sqrt{n}} \times \frac{1}{\lambda_p\sqrt{n}}\right)$ , es decir  $o_P(1)$ . El último término es por tanto  $o_P(1)$ . Puede concluirse que

$$\sup_{x} D(x) \le \left( (b + o_P(1)) \|y\| + (b + o_P(1))^2 \right) o_P(1),$$

y de ahí que  $\sup_x D(x) = o_P(1)$ .

De la consistencia de A, B, C y D puede concluirse que,  $\sup_x \mathcal{E}(x) = o_P(1)$ . Entonces, para cualquier  $\eta > 0$ ,

$$P\left(\sup_{x} |f(y \not x) - \hat{f}(y \not x)| > \eta\right) \leq P\left(\sup_{x} |f(y \not x) - \hat{f}(y \not x)| > \eta, \sup_{x} \mathcal{E}(x) \le \frac{1}{2}\right) + P\left(\sup_{x} \mathcal{E}(x) > \frac{1}{2}\right)$$

$$\leq P\left(\sup_{x} \mathcal{E}(x)e^{1/2}\sup_{x} f(y \swarrow x) > \eta\right) + P\left(\sup_{x} \mathcal{E}(x) > \frac{1}{2}\right).$$

porque, del supuesto (A12),  $\sup_x f(y \not/ x)$  es finito. Finalmente, como  $\sup_x \mathcal{E}(x) = o_P(1)$ ,  $\lim_{n \to +\infty} P\left(\sup_x |f(y \not/ x) - \hat{f}(y \not/ x)| > \eta\right) = 0$  lo cual concluye la demostración.  $\Box$ 

# 4.3.5. Conclusión: Consistencia de $\hat{\gamma}(y)$

Para la consistencia del estimador DBIR se requieren los siguientes supuestos adicionales:

(A13) para todo  $y \in \mathcal{Y}, \int_{\mathbb{R}} f(y \swarrow x) x f_X(x) dx$  es finito;

(A14)  $\mathbb{E}(X)$  es finito.

El resultado final está dado en el siguiente teorema:

**Teorema 4.1.** Bajo los supuestos (A1)-(A14), se tiene que, para todo  $y \in \mathcal{Y}$  tal que  $f_Y(y) > 0$ ,

$$\lim_{n \to +\infty} \hat{\gamma}(y) =^{P} \gamma(y).$$

Antes de comenzar la demostración del teorema es necesario demostrar la siguiente preposición. Recuérdese que:

- $f_Y(y) = \int_{\mathbb{R}} f(y \swarrow x) f_X(x) dx$ ;
- $\hat{f}_Y(y) = \frac{1}{n} \sum_{i=1}^n \hat{f}(y \neq x_i);$

**Proposición 4.4.** Bajo los supuestos (A1)-(A12), para cualquier  $y \in \mathcal{Y}$ , se tiene:

$$\left|f_Y(y) - \hat{f}_Y(y)\right| = o_P(1),$$

y, con los supuestos adicionales (A13) y (A14),

$$\left|\frac{1}{n}\sum_{i=1}^{n}\hat{f}(y \nearrow x_{i})x_{i} - \int_{\mathbb{R}}f(y \nearrow x)xf_{X}(x)dx\right| = o_{P}\left(1\right).$$

Demostración:

Para cualquier  $y \in \mathcal{Y}$ ,

$$\left|\hat{f}_{Y}(y) - f_{Y}(y)\right| \leq \left|\hat{f}_{Y}(y) - \frac{1}{n}\sum_{i=1}^{n}f(y \nearrow x_{i})\right| + \left|\frac{1}{n}\sum_{i=1}^{n}f(y \cancel{x}_{i}) - f_{Y}(y)\right|$$

pero

• por la Proposición 4.3,

$$\left|\hat{f}_{Y}(y) - \frac{1}{n}\sum_{i=1}^{n} f(y \nearrow x_{i})\right| \leq \sup_{x} \left|f(y \nearrow x) - \hat{f}(y \nearrow x)\right| = o_{P}(1);$$

• el supuesto (A12) garantiza que, para todo  $y \in \mathcal{Y}$ ,  $f_Y(y)$  es finito. Por tanto, por la Ley de los Grandes Números (LLN)

$$\lim_{n \to +\infty} \frac{1}{n} \sum_{i=1}^n f(y \not x_i) =^{ps} \mathbb{E}_X \left( f(y \not X) \right) = f_Y(y).$$

Estos dos argumentos completan la primera parte de la demostración.

La segunda parte se demuestra de la misma forma: para cualquier  $y \in \mathcal{Y}$ ,

$$\begin{aligned} \left| \frac{1}{n} \sum_{i=1}^{n} \hat{f}(y \swarrow x_{i}) x_{i} - \int_{\mathbb{R}} f(y \swarrow x) x f_{X}(x) dx \right| &\leq \left| \frac{1}{n} \sum_{i=1}^{n} \hat{f}(y \nearrow x_{i}) x_{i} - \frac{1}{n} \sum_{i=1}^{n} f(y \swarrow x_{i}) y_{i} \right| \\ &+ \left| \frac{1}{n} \sum_{i=1}^{n} f(y \swarrow x_{i}) x_{i} - \int_{\mathbb{R}} f(y \swarrow x) x f_{X}(x) dx \right| \end{aligned}$$

donde

• la primera parte del miembro derecho de la desigualdad previa está acotada por

$$\sup_{x} \left| f(y \swarrow x) - \hat{f}(y \swarrow x) \right| \times \frac{1}{n} \sum_{i} x_{i}.$$

Con el supuesto (A14),  $\lim_{n \to +\infty} \frac{1}{n} \sum_{i} x_i = \mathbb{E}(X)$  y por la Proposición 4.3,  $\sup_x \left| f(y \swarrow x) - \hat{f}(y \swarrow x) \right| = o_P(1)$ . Entonces,

$$\left|\frac{1}{n}\sum_{i=1}^{n}\hat{f}(y \nearrow x_{i})x_{i} - \frac{1}{n}\sum_{i=1}^{n}f(y \cancel{x}_{i})x_{i}\right| = o_{P}(1).$$

 la segunda parte del miembro derecho de la desigualdad converge a cero casi seguramente por LLN bajo el supuesto (A14). □

### Demostración del teorema:

Denótese por  $g(y) = \int_{\mathbb{R}} f(y \swarrow x) y f_X(x) dx$  y  $\hat{g}(y) = \frac{1}{n} \sum_{i=1}^n \hat{f}(y \nearrow x_i) x_i$ . Se tiene que

$$\begin{aligned} |\hat{\gamma}(y) - \gamma(y)| &= \left| \frac{\hat{g}(y)f_Y(y) - g(y)\hat{f}_Y(y)}{f_Y(y)\hat{f}_Y(y)} \right| \\ &\leq \left| \frac{1}{|f_Y(y)|} |g(y) - \hat{g}(y)| + \left| \frac{\hat{g}(y)}{f_Y(y)\hat{f}_Y(y)} \right| \left| f_Y(y) - \hat{f}_Y(y) \right|. \end{aligned}$$

Pero,

$$\begin{aligned} |\hat{g}(y)| &\leq ||\hat{g}(y) - g(y)| + g(y)| \\ \hat{f}_{Y}(y) &\geq ||f_{Y}(y) - |\hat{f}_{Y}(y) - f_{Y}(y)|| \end{aligned}$$

entonces,

$$\left|\hat{\gamma}(y) - \gamma(y)\right| \le \frac{|g(y) - \hat{g}(y)|}{f_Y(y)} + \frac{||\hat{g}(y) - g(y)| + |g(y)||}{\left|f_Y(y) - |\hat{f}_Y(y) - f_Y(y)|\right| f_Y(y)} \left|\hat{f}_Y(y) - f_Y(y)\right|.$$

Como  $f_Y(y)$  es finito (supuesto (A12)) y positivo, el primer término del miembro derecho de la desigualdad anterior es  $o_P(1)$  por la Proposición 4.4. También, por la Proposición 4.4, el segundo término tiende a  $\frac{|g(y)|}{f_Y(y)^2}o_P(1)$  el cual es trivialmente  $o_P(1)$ , ya que g(y) es finita para todo y (supuesto (A13)). Por tanto, puede concluirse que  $|\hat{\gamma}(y) - \gamma(y)| = o_P(1)$  lo cual completa la demostración.  $\Box$ 

# 4.4. Estudio de simulación

En esta sección se discuten la factibilidad y el comportamiento del método de regresión funcional no paramétrico descrito en la sección 4.2 a través de un estudio de simulación.

### 4.4.1. Generación de los datos

Los datos se simularon de la siguiente forma: los valores de la variable aleatoria real X fueron obtenidos de una distribución uniforme en el intervalo [0, 10]. Entonces,

Y fue generado por 4 modelos de regresión diferentes:

M1 
$$Y = Xe_1 + 2Xe_2 + 3Xe_5 + 4Xe_{10} + e$$
  
M2  $Y = (\exp(X)/\exp(10))e_1 + (X^2/100)e_2 + (X^3/1000)e_5 + \log(X+1)e_{10} + e$   
M3  $Y = \sin(X)e_1 + \log(X+1)e_5 + e$   
M4  $Y = \alpha \exp\left(\frac{X}{10}e_1\right) + e$ 

donde  $(e_i)_{i\geq 1}$  es la base trigonométrica en  $\mathcal{Y} = \mathcal{L}^2([0,1])$  (*i.e.*,  $e_{2k-1} = \sqrt{2}\cos(2\pi kt)$ , y  $e_{2k} = \sqrt{2}\sin(2\pi kt)$ ), y *e* un proceso Gaussiano independiente de *X* con media cero y operador de covarianza  $\Gamma_e = \sum_{j\geq 1} \frac{1}{j}e_j \otimes e_j$ . De forma más precisa,  $\epsilon$  se simuló usando un truncamiento de  $\Gamma_e$ ,  $\Gamma_e(s,t) \simeq \sum_{j=1}^q \frac{1}{j}e_j(t)e_j(s)$  con q = 500. Se simularon muestras de calibración y prueba de tamaños  $n_L = 300$  y  $n_T = 200$  respectivamente sobre la base de estos modelos.

Las Figuras 4.1 y 4.2 muestran ejemplos de realizaciones de Y obtenidas bajo el primer y tercer modelo, respectivamente, para tres valores diferentes de x. También se representan las funciones subyacentes (no ruidosas), F(x). Nótese que en el caso del modelo **M3**, los datos simulados tienen un alto nivel de ruido lo cual hace de la estimación una tarea difícil.



Figura 4.1: Modelo **M1**. Función verdadera F(x) (línea continua suave), datos simulados Y (línea gris rugosa) y aproximación de Y usando B-splines (linea negra rugosa) para tres valores diferentes de x.

Para aplicar el método DBIR, las funciones discretizadas Y se aproximaron por funciones continuas usando desarrollo en funciones bases. Específicamente, los datos se aproximaron usando una base B-spline de orden cuatro y 128 funciones bases. Las Figuras 4.1 y 4.2 muestran la comparación entre las funciones originales y sus aproximaciones B-spline.



Figura 4.2: Modelo **M3**. Función verdadera F(x) (línea continua suave), datos simulados Y (línea gris rugosa) y aproximación de Y usando B-splines (linea negra rugosa) para tres valores diferentes de x.

### 4.4.2. Resultados de la simulación

La media condicional  $\mu(x)$  se estimó usando un estimador de núcleo (como en la ecuación (4.3)). Dos hiperparámetros tuvieron que ajustarse en este enfoque: el parámetro ancho de banda para la estimación de  $\mu(x)$  y el número, p, de funciones propias involucradas en la Ecuación (4.4). Estos dos hiperparámetros se seleccionaron sobre la base de una validación cruzada con 10 particiones que minimizara un criterio de error cuadrático medio en la muestra de entrenamiento. Por simplicidad, se muestran los resultados de manera gráfica solo para los modelos **M1** (caso lineal) y **M3** (un caso no lineal). No obstante, los resultados para los cuatro modelos, así como la comparación con el método NWK, se muestra en la tabla 4.1, al final de la sección.

### Caso lineal M1

Esta sección muestra los resultados obtenidos en el primer modelo simulado, en el que la variable explicativa real X y la variable respuesta funcional Y están relacionadas de manera lineal en el modelo inverso de la Ecuación (3.1). La Figura 4.1 muestra que el nivel de ruido en los datos es mayor para valores pequeños de X.

La estimación de la media condicional usando estimadores de núcleos (Ecuación (4.3)) se muestra en la Figura 4.3. De manera más específica, se grafican los valores reales y los valores estimados de F(x)(t) para diferentes valores de x (arriba) y para diferentes valores de t (abajo). De igual manera, puede constatarse en esta figura, la linealidad del modelo inverso (Ecuación (3.1)). En general, las estimaciones



Figura 4.3: Modelo **M1**. Arriba: valores verdaderos (líneas discontinuas) y estimaciones (líneas continuas) de F(x) para varios valores de x. Abajo: valores verdaderos y estimaciones de  $F(\cdot)(t)$  para varios valores de t (abajo). Los puntos representan los datos simulados  $(x_i(t))_i$  en el conjunto de entrenamiento.

son buenas pero en algunos casos (e.g., abajo derecha) el nivel de ruido parece ser muy alto y la media verdadera (vista como función de x) no queda tan bien estimada como en otros casos.

La Figura 4.4 muestra la descomposición espectral de la estimación de la covarianza empírica de los residuos  $\Gamma$  (sección 4.2.2, paso 2) y los valores predichos de Xpor DBIR en los conjuntos de entrenamiento y de prueba. En particular, la comparación entre las funciones propias verdaderas y las estimadas se muestra en la Figura 4.4 (a-c) y la comparación entre los valores propios verdaderos y los estimados, en la Figura 4.4 (d). Los resultados de la predicción son muy buenos, tanto en el conjunto de entrenamiento como en el de prueba (Figura 4.4 (e-f)).



Figura 4.4: Modelo **M1.** (a-c): Funciones propias verdaderas (línea discontinua) y estimaciones (línea continua); (d): valores propios estimados vs. valores propios verdaderos y (e-f): valores predichos para X vs. valores verdaderos para el conjunto de entrenamiento y de prueba.

### Caso no lineal M3

Para el modelo no lineal **M3**, puede apreciarse en la Figura 4.2 que el nivel de ruido es mucho mayor que para el modelo **M1**. De forma análoga a como se hizo en la sección anterior, la Figura 4.5 compara la verdadera F(x)(t) con su estimación para varios valores de x (arriba) y para varios valores de t (abajo). Los resultados son muy satisfactorios dado el hecho de que los datos presentan altos niveles de ruido (lo cual se aprecia claramente en la parte inferior de esta figura).

La Figura 4.6 muestra los resultados de los pasos 2-3 del esquema de estimación: donde la descomposición espectral de la estimación de  $\Gamma$  se compara con los valores verdaderos; y los valores predichos para X por DBIR se comparan con los valores verdaderos, tanto en el conjunto de entrenamiento como en el de prueba. La estimación de la descomposición espectral es también satisfactoria a pesar del alto nivel de ruido. Las comparaciones de las predicciones en los conjuntos de entrenamiento y de prueba muestran que el método no sobreentrena los datos.

Finalmente, el comportamiento de DBIR para cada uno de los modelos se comparó



Figura 4.5: Modelo **M3**. Arriba: valores verdaderos (líneas discontinuas) y estimaciones (líneas continuas) de F(x) para varios valores de x. Abajo: valores verdaderos y estimaciones de  $F(\cdot)(t)$  para varios valores de t (abajo). Los puntos (abajo) son los datos simulados  $(x_i(t))_i$  en el conjunto de entrenamiento.

con el obtenido por el estimador NWK funcional con dos tipos de métricas dentro del núcleo: la norma  $L_2$  usual y la norma basada en la semi-métrica de PCA (ver [27] para mayor detalle sobre estos métodos). Los valores de RMSE resultantes se presentan en la Tabla 4.1.

Model	DBIR	NWK (pca)	$NWK \ (\mathcal{L}^2)$
M1	0.08	0.10	0.09
M2	1.47	1.60	1.77
M3	1.79	1.79	2.00
$\mathbf{M4}$	0.94	2.16	1.91

Tabla 4.1: RMSE para cada método en cada uno de los modelos.

Los resultados muestran que DBIR es una buena alternativa a métodos comunes como es el caso de NWK. De hecho, DBIR se comporta mejor que el método NWK en todos los casos considerados en este estudio de simulación que incluye modelos lineales (M1) y modelos no lineales (M2 - M4).



Figura 4.6: Modelo **M3**: (a-c): Funciones propias verdaderas (línea discontinua) y estimaciones (línea continua); (d): valores propios estimados vs. valores propios verdaderos y (e-f): valores predichos para X vs. valores verdaderos para el conjunto de entrenamiento y de prueba.

# 4.5. Resultados en el conjunto de datos "Tecator"

DBIR se probó también en el conjunto de datos Tecator, descrito en el Capítulo 2. Este problema es más complejo que el que se presentó en la sección 4.4 donde los datos fueron generados bajo los supuestos del método. En esta sección, el objetivo que se persigue con la aplicación de DBIR a Tecator no es optimizar la solución para la problemática que trae consigo el conjunto de datos (para esto debe realizarse un estudio más detallado de las diferentes variantes de pre-procesamiento, selección de hiperparámetros, normas, etc), sino ilustrar que la nueva propuesta se comporta razonablemente bien en un conjunto de datos real y complejo que puede desviarse de los supuestos de DBIR, comparada con otros enfoques alternativos, como es el caso de NWK.

Todo el conjunto de datos se dividió para formar dos conjuntos, el de entrenamiento y el de prueba, ambos prácticamente del mismo tamaño. Esta división se realizó de manera aleatoria y se repitió 100 veces, garantizando que cada conjunto tuviera datos con valores de contenido de grasa representativos de todo el rango de calibración.

La Tabla 4.2 reporta la media del MSE (y su desviación estándar) para ambos mé-

todos (DBIR y NWK) calculado en las 100 divisiones del conjunto de datos realizadas.

Model	DBIR	NWK (pca)	$NWK \ (\mathcal{L}^2)$
MSE	1.91(0.41)	9.1(2.1)	8.9(2.1)

Tabla 4.2: Resultados de predicción en el conjunto de datos "Tecator"

Los resultados obtenidos por DBIR son satisfactorios y mejores que los obtenidos por el método NWK, bajo el criterio de MSE, lo cual evidencia que DBIR puede resultar una alternativa interesante para la solución de problemas de calibración funcional. En el Capítulo 3 se vio que para este conjunto de datos la segunda derivada del espectro contiene información útil. En la literatura se reportan [27] los resultados obtenidos usando NWK sobre la base de una semi-métrica que involucra la segunda derivada; el valor de MSE obtenido en dicho estudio es de 3.5, el cual es aún mayor que el obtenido en esta experimentación por DBIR (sin hacer uso de información sobre las derivadas).
# CONCLUSIONES Y RECOMENDACIONES

Para concluir esta tesis se tiene que:

- 1. Se introdujo un nuevo enfoque predictivo no Bayesiano (frecuentista) para el problema de calibración, el cual se elaboró en detalles para la calibración univariada lineal bajo supuestos de Gaussianidad. Este enfoque proporciona una función de verosimilitud predictiva no Bayesiana para el parámetro de calibración desconocido  $x_0$  y un estimador puntual  $\hat{x}_P$  del mismo. Para este estimador se estudiaron las siguientes propiedades:
  - Finitud de los momentos.
  - Consistencia del estimador.
  - Desarrollo asintótico del estimador.
  - Desarrollo asintótico del Error Cuadrático Medio. En particular, se demostró que tiene MSE acotado hasta la aproximación de orden cero (con respecto al tamaño de muestra de entrenamiento n), como el estimador clásico  $\hat{x}_C$  y los estimadores  $\hat{x}_{SS}$ ,  $\hat{x}_N$  propuestos en [87] y [68], respectivamente.
- 2. Se realizó un estudio de simulación con tamaño de muestra moderado, en el que (apoyado también en el análisis asintótico) se muestra que  $\hat{x}_P$  es mejor que  $\hat{x}_C$  y  $\hat{x}_{SS}$  en todo el rango de posibles valores de  $x_0$  bajo el criterio de RMSE y D50. Para problemas de extrapolación (i.e., cuando  $x_0$  está fuera del rango de calibración),  $\hat{x}_P$  resultó ser mejor que el estimador inverso  $\hat{x}_I$  y tener un comportamiento muy similar a  $\hat{x}_N$ . Por tanto, podría concluirse que  $\hat{x}_P$  y  $\hat{x}_N$  resultan muy atractivos para la extrapolación. Para la interpolación (i.e., cuando  $x_0$  está dentro del rango de calibración),  $\hat{x}_P$  resulto de calibración.

como  $\hat{x}_C$  y  $\hat{x}_{SS}$ . Adicionalmente, se utilizó la distancia de Kullback-Leibler para una comparación de  $\hat{x}_P$  con el estimador de Naszódi,  $\hat{x}_N$ , debido a la similitud que mostraban ambos bajo los otros criterios usados, resultando  $\hat{x}_P$  mejor. Por tanto,  $\hat{x}_P$  parece ser un estimador puntual atractivo, especialmente para la extrapolación. El enfoque introducido tiene además varias características ventajosas que faltan en otros enfoques Bayesianos y no Bayesianos de calibración. En particular:

- No sólo proporciona un estimador puntual sino que también hace posible acceder a la verosimilitud a través de la función de verosimilitud predictiva no Bayesiana.
- Puede ser extendido a cualquier modelo estadístico paramétrico (regular) como los multivariados, no Gaussianos, no lineales e incluso funcionales.
- 3. Se aplicó el método de regresión no paramétrico Máquinas de vectores relevantes en la solución de problemas de calibración multivariada. RVMR fue aplicado a tres conjuntos de datos reales de espectroscopía, uno de ellos lineal y los otros no lineales. RVMR se comparó con otros métodos no paramétricos usados para la calibración multivariada: SVR y LSSVM y con el método lineal PLS. Los resultados obtenidos en los experimentos indican que RVMR parece ser una alternativa valiosa para la solución de problemas de calibración multivariada. Los modelos obtenidos con RVMR son altamente ralos, lo que los hace simples y precisos para la estimación de propiedades de interés.
- 4. Se introdujo un nuevo método de regresión no paramétrico funcional que consiste en la extensión del método de las Máquinas de vectores soporte a datos funcionales, en el marco de la aproximación de funciones en espacios de Hilbert con núcleos reproductores. El método propuesto se comparó con otros dos métodos de regresión no paramétricos funcionales, NWK y RBF sobre la base del conjunto de datos Tecator. FSVR mostró los mejores resultados bajo diversos criterios de evaluación, seguido por RBF. Quedó demostrada la capacidad de reducción de la dimensión (raleza) de FSVR a través de la obtención de modelos ralos. Se realizó además, una comparación con métodos de calibración multivariada como SVR, LSSVM y PLS, donde FSVR mostró los mejores resultados, lo que evidencia que tener en cuenta la naturaleza funcional de los datos es ventajoso. A pesar de las ventajas mencionadas, debe decirse que la estimación del modelo FSVR involucra tiempos computacionales ligeramente mayores y mayor

número de hiperparámetros que los métodos no paramétricos funcionales RBF y NWK.

Se derivó una clase de núcleos definidos no negativos apropiados a datos funcionales, como los núcleos  $\alpha$ -estables.

Debido a lo reciente que resulta esta temática, existen pocos resultados teóricos en cuanto a la consistencia y otras propiedades de estos tipos de estimadores. Esto es un aspecto que vale la pena tener en cuenta para futuras investigaciones.

5. Se desarrolló un nuevo enfoque de regresión funcional no paramétrica. El nuevo método llamado Regresión Inversa Basado en la Densidad se describió completamente para el espacio muestral  $\mathcal{Y} = L_2[a, b]$  y el supuesto de Gaussianidad para la ley condicional  $P(\cdot / X)$ . Dos características atractivas de DBIR, además de hacer uso explícito de la densidad del modelo, son que no requiere supuestos tan restrictivos como los de FIR y es computacionalmente simple.

El estudio de simulación desarrollado mostró resultados satisfactorios en situaciones de linealidad y no linealidad. DBIR fue aplicado al conjunto de datos reales Tecator y comparado con el método de regresión funcional no paramétrico NWK mostrando buenos resultados. La nueva propuesta es una solución promisoria para la calibración funcional.

#### Recomendaciones

Se recomienda:

- 1. La extensión del enfoque predictivo no Bayesiano desarrollado para la calibración univariada a modelos de calibración con datos multivariados y funcionales, no lineales y con supuestos no Gaussianos.
- 2. La aplicación de la clase de funciones núcleos  $\alpha$ -estables en la solución de problemas reales, así como la investigación en la definición de nuevos núcleos apropiados a este tipo de datos, problema que continúa abierto en el marco de la aproximación de funciones en RKHS, debido a que las propiedades de suavidad del estimador de la función de regresión, como un funcional en el espacio de funciones, no ha sido aún definida.
- 3. La extensión del método DBIR a otros espacios muestrales y familias de distribuciones.

### Bibliografía

- [1] AITCHISON, J. and DUNSMORE, I. R. (2002). *Statistical Prediction Analysis*. Cambridge University Press, Londres.
- [2] ALI, M. A. and SINGH, N. (1981). An alternative estimator in inverse regression, Journal of Statistical Computation and Simulation, 14, 1–15.
- [3] BELOUSOV, A. I., VERSAKOV, S. A., and FRESE, J. V. (2002). Application aspects of supports vector machines, *Journal of Chemometrics*, 16, 482–489.
- [4] BERKSON, J. (1969). Estimation of a linear function for a calibration line; consideration of a recent proposal, *Technometrics*, 11, 649–660.
- [5] BIAU, G., CÉROU, F., and GUYADER, A. (2010). Rates of convergence of the functional k-nearest neighbor estimate, *IEEE Transactions on Information Theory*, 56, 2034–2040.
- [6] BISHOP, C. M. (1995). Neural Networks for Pattern Recognition. Oxford University Press, Oxford.
- [7] BORGGAARD, C. and THODBERG, H. H. (1992). Optimal minimal neural interpretation of spectra, *Analytical Chemistry*, 64, 545–551.
- [8] BOSQ, D. (1991). Modelization, non-parametric estimation and prediction for continuous time processes, 335 of ASI Series, 509–529. NATO.
- [9] BOUSQUET, O. and ELISSEEF, A. (2002). Stability and generalization, *Journal of Machine Learning Research*, 1.
- [10] BROWN, P. J. (1982). Multivariate calibration (with discussion), Journal of the Royal Statistical Society, Series B, 44, 287–321.
- [11] BROWN, P. J. (1993). Measurement, Regression, and Calibration. Clarendon Press, Oxford.

- [12] CARDOT, H., CRAMBES, C., KNEIP, A., and SARDA, P. (2007). Smoothing spline estimators in functional linear regression with errors in variables, *Computional Statistics and Data Analysis*, 51, 4832–4848.
- [13] CARDOT, H., FERRATY, F., and SARDA, P. (1999). Functional linear model, Statistics and Probability Letter, 45, 11–22.
- [14] CARDOT, H., FERRATY, F., and SARDA, P. (1999). Functional linear model, Statistics and Probability Letters, 45, 11–22.
- [15] CARDOT, H., FERRATY, F., and SARDA, P. (2003). Spline estimators for the functional linear model, *Statistica Sinica*, 13, 571–591.
- [16] CENTNER, V., MASSART, D. L., and DE JONG, S. (1998). Inverse calibration predicts better than classical calibration, *Fresenius. J. Anal. Chem.*, 361, 2–9.
- [17] CHOW, S. and SHAO, J. (1990). On the difference between the classical and the inverse methods of calibration, *Journal of the Royal Statistical Society, C-Appl.*, 39, 219–228.
- [18] CRAMBES, C., KNEIP, A., and SARDA, P. (2008). Smoothing splines estimators for functional linear regression, *The Annals of Statistics*, 37, 35–72.
- [19] CRISTIANINI, N. and TAYLOR, J. S. (2000). An introduction to support vector machines and other kernel-based learning methods. Cambridge University Press, GB.
- [20] DAGO, A., TALAVERA, I., FERNÁNDEZ, R., HERNÁNDEZ, N., and LAZA, M. (2008). Predicting percent of crystallinity in FCC catalysts by FT- MIR and PLS, *Journal of Chemometrics*, 22, 661–665.
- [21] DAUXOIS, J., FERRÉ, L., and YAO, A. (2001). Un modèle semi-paramétrique pour variable aléatoire hilbertienne, *Compte Rendu de l'Académie des Sciences de Paris, Série I (Mathématique)*, 327(I), 6947–952.
- [22] DOLORES, M., DAGO, A., TALAVERA, I., HERNÁNDEZ, N., LAZA, M., and PORRO, D. (2009). A multivariate calibration approach for determination of petroleum hydrocarbons in water by means of IR spectroscopy, *Journal of Chemometrics*, 24, 444–447.

- [23] DYRBY, M., ENGELSEN, S. B., NØRGAARD, L., BRUHN, M., and NIELSEN, L. L. (2002). Chemometric quantitation of the active substance in a pharmacentrical tablet using near infrared (NIR) transmittance and NIR FT Raman spectra, *Applied Spectroscopy*, 56, 579–585.
- [24] EISENHART, C. (1939). The interpretation of certain regression methods and their use in biological and industrial research, Ann. Math. Statist., 10, 162–186.
- [25] EUGENIOU, T., PONTIL, M., and POGGIO, T. (2000). Regularization networks and support vector machines in Advances in Large Margin Classifiers (A. J. SMOLA, P. L. B. and SCHÖLKOPF, B., eds.), 171–203, The MIT Press, Cambridge.
- [26] FDA functions. Available at, *ftp://ego.psych.mcgill.ca/pub/ramsay/FDAfuns*.
- [27] FERRATY, F. and VIEU, P. (2006). Nonparametric Functional Data Analysis: Theory and Practice (Springer Series in Statistics). Springer-Verlag, Nueva York.
- [28] FERRÉ, L. (1998). Determining the dimension in sliced inverse regression and related methods, *Journal of the American Statistical Association*, 93, 132–140.
- [29] FERRÉ, L. and VILLA, N. (2006). Multi-layer perceptron with functional inputs
  : an inverse regression approach, *Scandinavian Journal of Statistics*, 33, 807–823.
- [30] FERRÉ, L. and YAO, A. (2003). Functional sliced inverse regression analysis, Statistics, 37, 475–488.
- [31] FERRÉ, L. and YAO, A. (2005). Smoothed functional inverse regression, Statistica Sinica, 15, no. 3, 665–683.
- [32] FRANK, I. E. and FRIEDMAN, J. H. (1993). A statistical view of some chemometrics regression tools, *Technometrics*, 35, 109–147.
- [33] GELADI, P., MACDOUGALL, D., and MARTENS, H. (1985). Linearization and scatter-correction for near-infrared reflectance spectra of meat, *Applied Spec*troscopy, 39, 491–500.
- [34] GIROSI, F., JONES, M., and POGGIO, T. (1995). Regularization theory and neural networks architectures, *Neural Computation*, 7.

- [35] HAGWOOD, C. (1992). The calibration problem as an ill-posed inverse problem, Journal of Statistical Planning and Inference, 31, 179–185.
- [36] HALPERIN, M. (1970). On inverse estimation in linear regression, *Technome-trics*, 12, 727–736.
- [37] HARRIS, I. R. (1989). Predictive fit for natural exponential families, *Biometri-ka*, 76, 675–684.
- [38] HASTIE, T. and MALLOWS, C. (1993). A discussion of a statistical view of some chemometrics regression tools by I. E. Frank and J. H. Friedman, *Technometrics*, 35, 140–143.
- [39] HELLAND, I. S. (1988). On the structured of partial least square regression, Communications in Statistics, 17, 581–607.
- [40] HELLAND, I. S. (2001). Some theoretical aspects of partial least squares regression, Chemometrics and Intelligent Laboratory Systems, 58, 97–107.
- [41] HERNÁNDEZ, N., BISCAY, R. J., and TALAVERA, I. (2007). Support vector regression methods for functional data, *Lecture Notes in Computer Science*, 4756, 564–573.
- [42] HERNÁNDEZ, N., BISCAY, R. J., and TALAVERA, I. (2010). A non bayesian predictive approach for statistical calibration, *Journal of Statistical Computation and Simulation*, Aceptado.
- [43] HERNÁNDEZ, N., BISCAY, R. J., VILLA, N., and TALAVERA, I. (2010). A functional density-based nonparametric approach for statistical calibration, *Lecture Notes in Computer Science*, Aceptado.
- [44] HERNÁNDEZ, N., BISCAY, R. J., VILLA, N., and TALAVERA, I. (2010). A simulation study of functional density-based inverse regression, *Investigación Operacional*, Enviado.
- [45] HERNÁNDEZ, N., KIRALJ, R., FERREIRA, M., and TALAVERA, I. (2009). Critical comparative analysis, validation and interpretation of SVM and PLS regression models in a QSAR study on HIV-1 protease inhibitors, *Chemometrics* and Intelligent Laboratory Systems, 98, 65–77.

- [46] HERNÁNDEZ, N., TALAVERA, I., BISCAY, R. J., PORRO, D., and FERREIRA, M. M. C. (2009). Support vector regression for functional data in multivariate calibration problems, *Analytica Chimica Acta*, 642, 110–116.
- [47] HERNÁNDEZ, N., TALAVERA, I., DAGO, A., BISCAY, R. J., FERREIRA, M. M. C., and PORRO, D. (2008). Relevance vector machines for multivariate calibration purposes, *Journal of Chemometrics*, 22, 686–694.
- [48] HERNÁNDEZ, N., TALAVERA, I., and PORRO, D. (2010). Regularized sliced inverse regression for determination of percent of crystallinity in FCC catalysts, *Journal of Chemometrics*, 24, 448–453.
- [49] HOERL, A. E. and KENNARD, R. W. (1970). Ridge regression: applications to nonorthogonal problems, *Technometrics*, 12, 69–82.
- [50] HOERL, A. E. and KENNARD, R. W. (1970). Ridge regression: biased estimation for nonorthogonal problems, *Technometrics*, 12, 55–67.
- [51] HÜLS, R. (1996). Der mittlere quadratische Fehler verschiedener Kalibrationssch Tesis de Maestría, Universidad de Dortmund.
- [52] KRUTCHKOFF, R. G. (1967). Classical and inverse regression methods of calibration, *Technometrics*, 9, 425–439.
- [53] KRUTCHKOFF, R. G. (1969). Classical and inverse regression methods of calibration in extrapolation, *Technometrics*, 11, 605–608.
- [54] KSHIRSAGAR, A. M. (1972). *Multivariate analysis*. Marcel Dekker, Nueva York.
- [55] KULLBACK, S. (1978). Information Theory and Statistics. Dover Publications, Inc.
- [56] LALOË, T. (2008). A k-nearest neighbor approach for functional regression, Statistics and Probability Letters, 78, 1189–1193.
- [57] LI, K. (1991). Sliced inverse regression for dimension reduction, Journal of the American Statistical Association, 86, 316–327.
- [58] LS-SVMLAB: a matlab/c toolbox for Least Squares Support Vector Machines. Available at, http://www.esat.kuleuven.ac.be/sista/lssvmlab/.

- [59] LWIN, T. and MARITZ, J. S. (1980). A note on the problem of statistical calibration, Journal of the Royal Statistical Society. Series C, 29, 135–141.
- [60] LWIN, T. and MARITZ, J. S. (1982). An analysis of the linear calibration controversy from the perspective of compound estimation, *Technometrics*, 24, 235–242.
- [61] MARTENS, H. and NAES, T. (1989). Multivariate calibration. Wiley, Chichester.
- [62] MARX, B. D. and EILERS, P. H. (1999). Generalized linear regression on sampled signals and curves: a P-spline approach, *Technometrics*, 41, 1–13.
- [63] MASSART, D. L., VANDEGINSTE, B. G. M., BUYDENS, L., JONG, S. D., LEWI, P. J., and SMEYERS-VERBEKE, J. (1997). Handbook of Chemometrics and Qualimetrics: Part B. Elsevier Science B. V., Holanda.
- [64] MASSY, W. F. (1965). Principal components regression in exploratory statistical research, Journal of the American Statistical Association, 60, 234–246.
- [65] MATLAB R2009a. The mathworks, Inc., Natick, MA.
- [66] MATLAB toolbox for kernel methods: the spider. Available at, http://www.kyb.tuebingen.mpg.de/bs/people/spider/.
- [67] NADARAYA, E. A. (1964). On estimating regression, Theory of Probability and its Applications, 9, 141–142.
- [68] NASZÓDI, L. J. (1978). Elimination of the bias in the course of calibration, *Technometrics*, 20, 201–205.
- [69] OSBORNE, C. (1991). Statistical calibration: A review, International Statistical Review, 59, 309–336.
- [70] PAWITAN, Y. (2001). In all likelihood. Statistical modelling and inference using likelihood. Clarendon Press, Oxford.
- [71] PERNG, S. K. (1987). A note on the inverse estimator for the linear calibration problem, *Commun. Stat-Theor. M.*, 16, 1743–1747.
- [72] PORRO, D., HERNÁNDEZ, N., TALAVERA, I., NÚÑEZ, O., DAGO, A., and BISCAY, R. J. (2008). Performance evaluation of relevance vector machines as a nonlinear regression method in real-world chemical spectroscopy data in

19th International Conference on Pattern Recognition (ICPR), IEEE Computer Society, 1–4.

- [73] PREDA, C. (2007). Regression models for functional data by reproducing kernel Hilbert space methods, *Journal of Statistical Planning and Inference*, 137, 829– 840.
- [74] PREDA, C. and SAPORTA, G. (2005). PLS regression on stochastic processes, Computional Statistics and Data Analysis, 48, 149–158.
- [75] QUIMIOMETRIX: Sistema de herramientas quimiométricas para el procesamiento, clasificación y calibración de datas químicas espectrales, *CENATAV*.
- [76] RAMSAY, J. and SILVERMAN, B. (2005). Functional Data Analysis. Springer, Nueva York, segunda ed.
- [77] RAO, B. (1983). Non Parametric Functional Estimation. Academic Press, Orlando, Florida, EU.
- [78] ROSIPAL, R. and TREJO, L. J. (2001). Kernel partial least squares regression in reproducing kernel Hilbert space, *Journal of Machine Learning Research*, 2, 97–123.
- [79] ROSSI, F. and CONAN-GUEZ, B. (2005). Functional multi-layer perceptron: a nonlinear tool for functional data anlysis, *Neural Networks*, 18, 45–60.
- [80] ROSSI, F., DELANNAY, N., CONAN-GUEZ, B., and VERLEYSEN, M. (2005). Representation of functional data in neural networks, *Neurocomputing*, 64, 183–210.
- [81] SAMORODNITSKY, G. and TAQQU, M. S. (1994). Stable Non-Gaussian Random Processes. Chapman & Hall, Nueva York.
- [82] SCHÖLKOPF, B. and SMOLA, A. (2002). *Learning with Kernels*. The MIT Press, Cambridge.
- [83] SERFLING, R. J. (1980). Approximation Theorems of Mathematical Statistics. John Wiley & Sons, Nueva York.
- [84] SHUKLA, G. (1972). On the problem of calibration, *Technometrics*, 14, 547–553.

- [85] SMOLA, A., SCHOLKOPF, B., and MULLER, K. (1998). The connection between regularization operators and support vector kernels, *Neural Networks*, 11, 637– 649.
- [86] SPROTT, D. A. (2000). Statistical inference in science. Springer-Verlag, Nueva York.
- [87] SRIVASTAVA, V. K. and SINGH, N. (1989). Small disturbance asymptotic theory for linear calibration estimators, *Technometrics*, 31, 373–378.
- [88] STONE, M. and BROOKS, R. J. (1990). Continuum regression: cross-validated sequentially constructed prediction embracing ordinary least squares, partial least squares and principal components regression, *Journal of the Royal Statistical Society, Series B*, 52, 237–269.
- [89] SUNDBERG, R. (1993). Continuum regression and ridge regression, Journal of the Royal Statistical Society, Series B, 55, 653–659.
- [90] SUYKENS, J., GESTEL, T. V., BRANMASTER, J. D., MOOR, B. D., and VAN-DEWALLE, J. (2002). Least Squares Support Vector Machines. World Scientific; Singapur.
- [91] TABLET Data set available at, http://www.models.kvl.dk.research/data/Tablets/.
- [92] TECATOR Data set available at, http://lib.stat.cmu.edu/datasets/tecator.
- [93] TELLINGHUISEN, J. (2000). Inverse vs. classical calibration for small data sets, Fresenius. J. Anal. Chem., 368, 585–588.
- [94] THISSEN, U., PEPERS, M., ÜSTUN, B., MELSSEN, W., and BUYDENS, L. (2004). Comparing support vector machines to PLS for spectral regression applications, *Chemometrics and Intelligent Laboratory Systems*, 73, 169–179.
- [95] THISSEN, U., ÜSTUN, B., MELSSEN, W., and BUYDENS, L. (2004). Multivariate calibration with least squares support vector machines, *Analytica Chemistry*, 76, 3099–3105.
- [96] TIPPING, M. E. (2000). The relevance vector machines in Advances in Neural Information Processing Systems (S. A. SOLLA, T. K. L. and MULLER, K.-R., eds.), 12, 652–658, MIT Press.

- [97] TIPPING, M. E. (2001). Sparse bayesian learning and the relevance vector machine, Journal of Machine Learning Research, 1, 211–244.
- [98] TIPPING, M. E. (2001). SparseBayes: a Matlab implementation of Sparse Bayesian Learning, Available at http://research.microsoft.com/mlp/rvm/.
- [99] TURIEL, T. P., HAHN, G. J., and TUCKER, W. T. (1982). New simulations results for the calibration and inverse median estimation problems, *Commun. Stat-Simul. C.*, 11, 677–713.
- [100] VAPNIK, V.Ñ. (1998). Statistical Learning Theory. John Wiley & Sons, Nueva York.
- [101] WAHBA, G. (1990). Spline Models for Observational Data. SIAM, Filadelfia.
- [102] WALCZAK, B. and MASSART, D. L. (1996). The radial basis functions partial least squares approach as a flexible non-linear regression techniques, *Analytica Chimica Acta*, 331, 177–185.
- [103] WALTERS, F. H. and RIZZUTO, G. T. (1988). The calibration problem in statistics and its application to chemistry, *Anal. Lett.*, 21, 2069–2076.
- [104] WILLIAMS, E. J. (1969). A note on regression methods in calibration, *Techno-metrics*, 11, 189–192.
- [105] WILLIAMS, E. J. (1969). Regression methods in calibration problems., Bull. Int. Statist. Inst., 43, 17–28.
- [106] WOLD, H. (1984). PLS Regression in *Encyclopaedia of Statistical Sciences* (JOHNSON, N. L. and KOTZ, S., eds.), 6, 581–591, John Wiley, Nueva York.
- [107] WOLD, S., MARTENS, H., and WOLD, H. (1983). The multivariate calibration problem in chemistry solved by PLS in *Matrix pencils* (RUHE, A. and KAGS-TROM, B., eds.), Lecture notes in mathematics, Springer, Heidelberg, 286–293.
- [108] WOLD, S., RUHE, H., WOLD, H., and DUNN, W. J. (1984). The collinearity problem in linear regression. the partial least squares (PLS) approach to generalized inverse, SIAM Journal of Scientific and Statistical Computations, 5, 735–743.

- [109] WÜLFERT, F., KOK, W. T., and SMILDE, A. K. (1998). Influence of temperature on vibrational spectra and consequences for the predictive ability of multivariate models, *Analytical Chemistry*, 70, 1761–1767.
- [110] YAO, A. (2001). Un modèle semi-paramérique pour variables fonctionnelles : la régression inverse fonctionnelle. Tesis de Doctorado, Universidad de Toulouse III, Francia.

## ANEXO A

#### Términos químicos

#### Absorbancia:

En espectroscopía, la absorbancia o absorbencia (A) representa la cantidad de luz que absorbe una muestra. Está definida como:

$$A_{\lambda} = -\log_{10}\left(\frac{I}{I_0}\right),\,$$

donde I es la intensidad de la luz a la longitud de onda  $\lambda$  pasada por la muestra (intensidad de la luz transmitida) e  $I_0$  la intensidad de la luz que se hace incidir a la muestra (intensidad de la luz incidente).

Las medidas de absorbancia son frecuentemente usadas en química analítica ya que se cumple la llamada *Ley de Lambert-Beer* que establece que la absorbancia es proporcional al grosor de una muestra y a la concentración de la sustancia en esta.

#### Espectro:

Se denomina espectro electromagnético a la distribución energética del conjunto de las ondas electromagnéticas. Referido a un objeto se denomina *espectro electromagnético* o simplemente *espectro* a la radiación electromagnética que emite (espectro de emisión) o absorbe (espectro de absorción) una sustancia. Dicha radiación sirve para identificar la sustancia de manera análoga a una huella dactilar. Los espectros se pueden observar mediante espectroscopios que, además de permitir observar el espectro, permiten realizar medidas sobre este, como la longitud de onda, la frecuencia, y la intensidad de la radiación. Fracción molar:

La fracción molar es una unidad química que expresa la concentración de soluto en una disolución y la proporción en que se encuentran los moles de soluto con respecto a los moles totales de disolución, que se calculan sumando los moles de soluto(s) y de disolvente.

### ANEXO B

#### Aspectos computacionales

El estudio del comportamiento de los diferentes métodos, así como las experimentaciones llevadas a cabo en la tesis se realizaron en MatLab [65]. Se creó una plataforma experimental que contiene la implementación de todos los métodos desarrollados en la tesis: el método predictivo no Bayesiano propuesto para la calibración univariada y los dos métodos para datos funcionales, FSVR y DBIR. La plataforma incluye además la implementación de algunos de los métodos usados para las comparaciones (SVR, LSSVM, RVMR, RBF, NWK). El desarrollo de la plataforma experimental estuvo apoyado en algunas librerías existentes como son: la SPIDER [66], que contiene la implementación LIBSVM para el método SVR; la librería LS-SVMLAB [58], que implementa el método LSSVM; la librería SparseBayes [98], que implementa el método RVMR y la librería MFDA [26] para el manejo de datos funcionales. También se implementaron los algoritmos de la validación cruzada utilizados para el ajuste de los hiperparámetros. El siguiente esquema resume la plataforma de experimentación desarrollada en Matlab en el marco de esta tesis.



Los métodos propuestos en el marco de esta tesis fueron implementados, además, en lenguaje C, e incorporados al "Sistema de Herramientas Quimiométricas para el procesamiento, clasificación y calibración de datos químicos espectrales: Quimiometrix" [75], desarrollado en el Centro de Aplicaciones de Tecnologías de Avanzada (CE-NATAV). Este sistema se encuentra hace dos años en explotación en varios centros de investigación del país, lo cual garantiza que los métodos propuestos se comiencen a usar en la solución de problemas reales.

RT\_046, enero 2012 Aprobado por el Consejo Científico CENATAV Derechos Reservados © CENATAV 2012 **Editor:** Lic. Lucía González Bayona **Diseño de Portada:** Di. Alejandro Pérez Abraham RNPS No. 2142 ISSN 2072-6287 **Indicaciones para los Autores:** Seguir la plantilla que aparece en www.cenatav.co.cu C E N A T A V 7ma. No. 21812 e/218 y 222, Rpto. Siboney, Playa; La Habana. Cuba. C.P. 12200 *Impreso en Cuba* 

