UNIVERSIDAD NACIONAL DEL LITORAL

DOCTORADO EN INGENIERÍA

Nuevos métodos basados en núcleos para la representación eficiente de datos bajo suficiencia estadística

Diego Isaías Ibañez

FICH
FACULTAD DE INGENIERÍA Y CIENCIAS HÍDRICAS

INTEC
INSTITUTO DE DESARROLLO TECNOLÓGICO PARA LA INDUSTRIA QUÍMICA

CIMEC
CENTRO DE INVESTIGACIÓN DE MÉTODOS COMPUTACIONALES

sinc(i)
INSTITUTO DE INVESTIGACIÓN EN SEÑALES, SISTEMAS E INTELIGENCIA COMPUTACIONAL

Tesis de Doctorado **2023**

UNIVERSIDAD NACIONAL DEL LITORAL

Facultad de Ingeniería y Ciencias Hídricas

Instituto de Desarrollo Tecnológico para la Industria Química

Centro de Investigación de Métodos Computacionales

Instituto de Investigación en Señales, Sistemas e Inteligencia Computacional

NUEVOS MÉTODOS BASADOS EN NÚCLEOS PARA LA REPRESENTACIÓN EFICIENTE DE DATOS BAJO SUFICIENCIA ESTADÍSTICA

Diego Isaías Ibañez

Tesis remitida al Comité Académico del Doctorado
como parte de los requisitos para la obtención del grado de
DOCTOR EN INGENIERÍA
Mención Inteligencia Computacional, Señales y Sistemas
de la
UNIVERSIDAD NACIONAL DEL LITORAL

2023

Comisión de Posgrado, Facultad de Ingeniería y Ciencias Hídricas,

Ciudad Universitaria, Paraje “El Pozo”, S3000, Santa Fe, Argentina.

UNIVERSIDAD NACIONAL DEL LITORAL

Facultad de Ingeniería y Ciencias Hídricas

Instituto de Desarrollo Tecnológico para la Industria Química

Centro de Investigación de Métodos Computacionales

Instituto de Investigación en Señales, Sistemas e Inteligencia Computacional

NUEVOS MÉTODOS BASADOS EN NÚCLEOS PARA LA REPRESENTACIÓN EFICIENTE DE DATOS BAJO SUFICIENCIA ESTADÍSTICA

Diego Isaías Ibañez

Lugar de trabajo:
Departamento de Matemática
Facultad de Ingeniería Química
Universidad Nacional del Litoral

Director:
Dr. Diego Tomassi — Biofortis SAS

Codirectora:
Dra. Liliana Forzani — CONICET - UNL

Jurado evaluador:
Dr. Gastón Schlottahuer — CONICET - IBB
Dra. Ana Georgina Flesia — CONICET - FAMAF
Dra. Daniela Rodríguez — CONICET - UBA

2023

UNIVERSIDAD NACIONAL DEL LITORAL
Facultad de Ingeniería y Ciencias Hídricas

Santa Fe, 14 de diciembre de 2023.

Como miembros del Jurado Evaluador de la Tesis de Doctorado en Ingeniería titulada ***"Nuevos métodos basados en núcleos para la representación eficiente de datos bajo suficiencia estadística"***, desarrollada por el Lic. Diego Isaías IBÁÑEZ, en el marco de la Mención "Inteligencia Computacional, Señales y Sistemas", certificamos que hemos evaluado la Tesis y recomendamos que sea aceptada como parte de los requisitos para la obtención del título de Doctor en Ingeniería.

La aprobación final de esta disertación estará condicionada a la presentación de dos copias encuadernadas de la versión final de la Tesis ante el Comité Académico del Doctorado en Ingeniería.

Dr. Gastón Schlottahuer	Dra. Ana Georgina Flesia	Dra. Daniela Rodríguez

Santa Fe, 14 de Diciembre de 2023.

Certifico haber leído la Tesis, preparada bajo mi dirección en el marco de la Mención "Inteligencia Computacional, Señales y Sistemas" y recomiendo que sea aceptada como parte de los requisitos para la obtención del título de Doctor en Ingeniería.

Dra. Liliana Forzani Codirector de Tesis	Dr. Diego Tomassi Director de Tesis

FACULTAD DE INGENIERÍA Y CIENCIAS HÍDRICAS * SANTA FE *
Dr. JOSÉ LUIS MACOR
SECRETARIO DE POSGRADO
Facultad de Ingeniería y Cs. Hídricas

Universidad Nacional del Litoral
Facultad de Ingeniería y
Ciencias Hídricas

Secretaría de Posgrado

Ciudad Universitaria
C.C. 217
Ruta Nacional Nº 168 – Km. 472,4
(3000) Santa Fe
Tel: (54) (0342) 4575 229
Fax: (54) (0342) 4575 224
E-mail: posgrado@fich.unl.edu.ar

Declaración legal del autor

Esta tesis ha sido remitida como parte de los requisitos para la obtención del grado académico de Doctor en Ingeniería, mención Inteligencia Computacional, Señales y Sistemas, ante la Universidad Nacional del Litoral y ha sido depositada en la Biblioteca de la Facultad de Ingeniería y Ciencias Hídricas para que esté a disposición de sus lectores bajo las condiciones estipuladas por el reglamento de la mencionada Biblioteca.

Citaciones breves de esta tesis son permitidas sin la necesidad de un permiso especial, en la suposición de que la fuente sea correctamente citada. Solicitudes de permiso para la citación extendida o para la reproducción parcial o total de este manuscrito serán concebidos por el portador legal del derecho de propiedad intelectual de la obra.

A Isabella y Valentina, por dar luz a mi vida.

Agradecimientos

A mis directores Diego y Liliana, por compartirme todo su conocimiento, por su paciencia y por ayudarme a tomar decisiones acertadas. Además de cumplir su rol profesional, me han demostrado su calidez humana, brindándome confianza y alentándome día a día.

A los miembros del jurado, por su tiempo y dedicación en la evaluación de esta tesis. Sus comentarios han sido de gran valor y muy enriquecedores en esta etapa final.

Al CONICET, por haber financiado mi estudio. A la FICH, por admitirme en su doctorado. A la FIQ, por haberme dado un lugar de trabajo cómodo y reconfortante.

A mis compañeros de oficina y de estudio, gracias por las charlas, los consejos y los momentos compartidos. Me llevo grandes amigos de esta etapa en Santa Fe.

A los profesores que me han impartido su sabiduría a lo largo de todos estos años, gracias por haberme ayudado a crecer académicamente.

A mis amigos, por sus incesantes mensajes de afecto y por reunirnos cada vez que regresaba. Exequiel, compartir la experiencia hizo más llevadero el tiempo lejos de casa.

A mis hermanos Ana, Natalia, Javier, Yanina y Daniela, por estar siempre pendientes y dispuestos a tenderme sus manos, acompañándome en cada momento y aliviando la distancia. Mi más sincero agradecimiento para ellos, así como para mis cuñados y sobrinos.

A mis padres, Susana y Pedro, por permitirme ser quien soy y por apoyarme a lo largo de mi vida. Esto es por y para ustedes. Los amo con toda mi alma.

A mi esposa Paula, por creer en mí y apoyarme incondicionalmente, convencida de que lo lograríamos. No hubiese sido posible sin vos. De corazón, gracias.

Índice general

Índice de figuras

Índice de tablas

Resumen

En muchas aplicaciones en las que intentamos predecir una variable $Y \in \mathbb{R}$ a partir de un conjunto de variables predictoras $\mathbf{X} \in \mathbb{R}^p$, la reducción de dimensiones es una herramienta adecuada para ayudar a comprender los datos medidos y visualizar las relaciones existentes entre las variables. Consiste en obtener representaciones de los datos en un espacio de dimensión menor que p, con el objetivo de facilitar el análisis exploratorio y el posterior tratamiento estadístico. En este marco, la *reducción suficiente de dimensiones* (SDR) es una metodología supervisada que intenta proporcionar una solución rigurosa al propósito de reducir $\mathbf{X}$ preservando la información sobre Y, utilizando el concepto de suficiencia estadística. La idea central es encontrar una transformación $\mathbf{R}(\mathbf{X})$ de dimensión $q \leq p$, de manera tal que el estudio de $Y|\mathbf{R}(\mathbf{X})$ sea equivalente al de $Y|\mathbf{X}$ pero con la ventaja de estar formulado en un espacio de dimensión posiblemente mucho menor.

La metodología de SDR para problemas de aprendizaje supervisado fue introducida en [Li, 1991] y formalizada luego en términos de distribuciones condicionales en [Cook, 1998]. Las propuestas iniciales [Li, 1991; Cook and Weisberg, 1991; Li and Wang, 2007; Bura and Cook, 2001] se basaron en funciones de momentos de la distribución de $\mathbf{X}|Y$, enfoque que se conoce como *regresión inversa*. Tales métodos proporcionaron típicamente transformaciones lineales de los predictores, con el objetivo de obtener el subespacio más pequeño capaz de preservar la información predictiva. Luego, dentro de este enfoque surgieron diferentes métodos basados en modelos de $\mathbf{X}|Y$, explotando frecuentemente los beneficios de las familias exponenciales. El resultado más general para estas familias se presentó en [Bura et al., 2016], donde hallaron de forma exhaustiva una SDR minimal de $\mathbf{X}$. Este resultado permitió ubicar la reducción de dimensiones supervisada de datos continuos y categóricos en el mismo marco conceptual. Además, se probó que la reducción

óptima no era lineal en los predictores sino en el estadístico suficiente de la familia exponencial elegida. En la práctica, una limitación de estos enfoques es que suele ser difícil evaluar si una suposición de modelado dada está respaldada por los datos.

En esta tesis buscamos ampliar la aplicabilidad de SDR a problemas reales, avanzando en dos direcciones vinculadas por la aplicación de métodos basados en núcleos en *espacios de Hilbert con núcleo reproductor* (RKHS) [Aronszajn, 1950].

En primer lugar, obtenemos SDR basadas en modelos para una amplia clase de distribuciones denominada *familias exponenciales basada en núcleos* (KEF) [Canu and Smola, 2006; Fukumizu, 2009]. Estos modelos probabilísticos comparten muchas propiedades con la familia exponencial clásica, pero pueden representar una gama mucho más amplia de distribuciones de probabilidad. Además, establecemos conexiones formales con clasificadores de vectores soporte (SVM) [Boser et al., 1992], que son relevantes para obtener reducciones eficientes desde el punto de vista computacional y que, a la vez, proveen un fundamento teórico que respalda el uso de SVM con fines de reducción y visualización.

En segundo lugar, abordamos el problema de reducción de dimensiones dentro del campo emergente de *aprender con información lateral*. Este campo se enfoca en un escenario predictivo especial donde, además de $\mathbf{X}$, existe otra variable $\mathbf{W}$ que contiene información sobre Y pero solo puede ser utilizada durante el proceso de estimación debido a diversas razones. En consecuencia, surge la temática de *reducción suficiente de dimensiones con información adicional*, la cual apenas ha sido abordada en la literatura. En este contexto, nuestra contribución consiste en una metodología general que aprovecha el potencial de los métodos basados en núcleos para manejar eficazmente una $\mathbf{W}$ de alta dimensionalidad.

Abstract

In many applications where we attempt to predict a variable $Y \in \mathbb{R}$ from a set of predictor variables $\mathbf{X} \in \mathbb{R}^p$, dimension reduction is an appropiate tool to aid understanding of the measured data and to visualize existing relationships between variables. It consists of obtaining representations of the data in a space of dimension less than p, in order to facilitate exploratory analysis and subsequent statistical treatment. In this framework, *sufficient dimension reduction* (SDR) is a supervising methodology that attempts to provide a rigorous solution for the objective of reducing $\mathbf{X}$ while preserving information about Y, by employing the concept of statistical sufficiency. The central idea is to find a transformation $\mathbf{R}(\mathbf{X})$ of dimension $q \leq p$, such that the study of $Y|\mathbf{R}(\mathbf{X})$ is equivalent to that of $Y|\mathbf{X}$ but with the advantage of being formulated in a space of possibly much smaller dimension.

SDR methodology for supervised learning problems was introduced in [Li, 1991] and later formalized in terms of conditional distributions in [Cook, 1998]. The initial proposals [Li, 1991; Cook and Weisberg, 1991; Li and Wang, 2007; Bura and Cook, 2001] were based on moment functions of $\mathbf{X}|Y$ distribution, an approach known as *inverse regression*. These methods typically provided linear transformations of the predictors, aiming to obtain the smallest subspace capable of preserving the predictive information. Within this approach, different methods emerged based on models of $\mathbf{X}|Y$, often exploiting the benefits of exponential families. The most general result on these families was presented in [Bura et al., 2016], where a minimal SDR of $\mathbf{X}$ was found exhaustively. Such result allowed to locate supervised dimension reduction of continuous and categorical data into the same conceptual framework. It was shown also that such optimal reduction was not linear in the predictors but in the sufficient statistic of the chosen exponential family. In

practice, one limitation of these approaches is that it is often difficult to assess if a given modeling assumption is supported by the data.

In this thesis we seek to extend the suitability of SDR to real problems, by advancing in two directions linked by the application of kernel methods in *reproducing kernel Hilbert spaces* (RKHS) [Aronszajn, 1950].

Firstly, we obtain model-based SDR for a broad class of distributions named *kernel exponential families* (KEF) [Canu and Smola, 2006; Fukumizu, 2009]. These probabilistic models share many properties with the classical exponential family, yet they can represent a much wider range of probability distributions. Furthermore, we establish formal connections with support vector classifiers (SVM) [Boser et al., 1992], that are relevant for obtaining computationally efficient reductions, while also providing a theoretical foundation that supports the use of SVM for the purposes of reduction and visualization.

Secondly, we address the dimension reduction problem within the emerging field of *learning with side information*. This field focuses on a special predictive scenario where, in addition to $\mathbf{X}$, there is another variable $\mathbf{W}$ that contains information about Y but can only be utilized during the estimation process due to various reasons. Consequently, the topic of *sufficient dimension reduction with additional information* arises, which has been hardly addressed in the literature. In this context, our contribution consists of a general methodology that leverages the potential of kernel-based methods to effectively handle a high-dimensional $\mathbf{W}$.

Glosario

Abreviaturas y siglas

v.a.	variable aleatoria	
s.d.p.	simétrica definida positiva	
RKHS	Espacio de Hilbert con núcleo reproductor *Reproducing Kernel Hilbert Space*	Def. 2.1
EF	Familia exponencial *Exponential Family*	Def. 2.5
KEF	Familia exponencial basada en núcleos *Kernel Exponential Family*	Def. 2.8
RKEF	Reducción en KEF restringida vía SVM	Def. 4.14
SDR	Reducción suficiente de dimensiones *Sufficient Dimension Reduction*	Def. 3.1
DRS	Subespacio de reducción suficiente *Dimension Reduction Subspace*	Def. 3.8
CS	Subespacio central *Central Subspace*	Def. 3.9
MLE	Estimación por máxima verosimilitud *Maximum Likelihood Estimation*	
PCA	Análisis de componentes principales *Principal Component Analysis*	
SIR	*Sliced Inverse Regression*	
SAVE	*Sliced Average Variance Estimation*	
PFC	*Principal Fitted Components*	
PLS	Mínimos cuadrados parciales *Partial Least Squares*	
EF-DR	Reducción de dimensiones para la familia exponencial *Exponential Family Dimension Reduction*	
COIR	*Covariance Operator Inverse Regression*	
LDA	Análisis discriminante lineal *Linear Discriminant Analysis*	
QDA	Análisis discriminante cuadrático *Quadratic Discriminant Analysis*	
KNN	K vecinos más cercanos *K Nearest Neighbors*	
SVM	Máquinas de vectores soporte *Support Vector Machines*	
MLP	Perceptrón multicapa *Multilayer Perceptron*	

Notaciones

Álgebra lineal

$\mathbf{I}_n$	Matriz identidad de orden n
$\mathbf{1}_n$	Matriz de unos de orden n
$\mathbf{A}^{\mathrm{T}}$	Traspuesta de la matriz $\mathbf{A}$
$\mathbf{A}^{-1}$	Inversa de la matriz $\mathbf{A}$
$\mathbf{A} > 0$	$\mathbf{A}$ es una matriz s.d.p.
$\operatorname{vec} \mathbf{A}$	Vectorización de la matriz $\mathbf{A}$
$\operatorname{vech} \mathbf{A}$	Vectorización de la parte triangular superior de la matriz simétrica $\mathbf{A}$
$\dim(\mathcal{S})$	Dimensión del subespacio $\mathcal{S}$
$\operatorname{span} A$	(a) Espacio generado por A, si A es una base; (b) Espacio columna de A, si A es una matriz
$\mathbf{P}_A$	Matriz de proyección ortogonal sobre: (a) A, si A es un subespacio; (b) $\operatorname{span} A$, si A es una base o una matriz
$\bigoplus_{i \in I} \mathcal{S}_i$	Suma directa de los subespacios $\mathcal{S}_i$

Análisis funcional

δ_{ij}	Función delta de Kronecker
$\langle \cdot, \cdot \rangle_{\mathbb{R}^n}$	Producto usual en $\mathbb{R}^n$
$\langle \cdot, \cdot \rangle_{\mathcal{H}}$	Producto interno en el espacio $\mathcal{H}$
$\| \cdot \|_{\mathcal{H}}$	Norma en $\mathcal{H}$ inducida por $\langle \cdot, \cdot \rangle_{\mathcal{H}}$
$C(\mathcal{X})$	Espacio de funciones continuas en $\mathcal{X}$
$C_0(\mathcal{X})$	Espacio de funciones continuas en $\mathcal{X}$ que se anulan en el infinito
$\| \cdot \|_p$	Norma p de vectores o funciones, según corresponda, con $1 \leq p \leq \infty$
$\| \cdot \|_u$	Norma uniforme o norma del supremo
$L^p(\mathcal{X})$	Espacio de funciones en $\mathcal{X}$ tales que $\|f\|_p < \infty$, $1 \leq p \leq \infty$
$\Sigma_{\mathbf{yx}}$, $\Sigma_{\mathbf{y}\|\mathbf{x}}$	Operadores de covarianza y covarianza condicional

Probabilidad y estadística

$A =_{\mathrm{D}} B$	A y B son v.a. idénticamente distribuidas
$A \perp\!\!\!\perp B$	A y B son v.a. independientes
$\mathbb{P}\{S\}$	Probabilidad del suceso S
$\mathbb{E}[X]$	Esperanza de X
$\mathrm{var}(X)$	Varianza o matriz de covarianza poblacional de X
$\mathrm{median}\, M$	Mediana de un conjunto de valores M
$\mathcal{D}_x$	Conjunto de datos de una v.a. X
$\boldsymbol{\mu},\ \boldsymbol{\Sigma}$	Esperanza y matriz de covarianza poblacional de una v.a. $\mathbf{X} \in \mathbb{R}^p$
$\bar{\mathbf{x}}, \mathbf{S}$	Media y matriz de covarianza muestrales de $\mathcal{D}_{\mathbf{x}}$
π_y	Probabilidad a priori de la clase $Y = y$
$\hat{\boldsymbol{\mu}},\ \hat{\boldsymbol{\Sigma}},\ \hat{\boldsymbol{\Delta}},\ \hat{\pi}_y$	MLE de $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$, $\boldsymbol{\Delta}$, π_y
$\mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Delta})$	Distribución Normal Multivariada en $\mathbb{R}^p$, con vector de medias $\boldsymbol{\mu}$ y matriz de covarianza $\boldsymbol{\Delta}$
$D_{\mathrm{KL}}(p, q)$	Divergencia de Kullback-Leibler entre p y q
$h(p, q)$	Distancia de Hellinger entre p y q
$J(p, q)$	Divergencia de Fisher entre p y q

Otros

$a \ll b$	a es mucho menor que b
$\mathbf{1}(\cdot)$	Función indicatriz

CAPÍTULO 1

Introducción

El Aprendizaje Automático (*Machine Learning*) es una disciplina cuya idea fundamental es que los algoritmos computacionales aprendan modelos predictivos a partir de un conjunto de datos que proporcionan ejemplos del fenómeno bajo estudio. Podría pensarse que el mejor escenario sería contar con una gran cantidad de variables que describan dicho fenómeno y también con una gran cantidad de casos u observaciones. Sin embargo, los grandes volúmenes de datos también pueden inducir complicaciones, tales como la cantidad de información redundante, la posible existencia de variables irrelevantes y la dificultad para contar con visualizaciones que ayuden a proponer modelos matemáticos con el objetivo de explicar las relaciones entre las variables.

Existen también numerosas situaciones en las cuales, a pesar de que las nuevas tecnologías permiten registrar un gran número de variables, no hay una posibilidad real de aumentar el tamaño muestral; en consecuencia, el número n de individuos observados es mucho menor que la cantidad p de características medidas ($n \ll p$). Este es típicamente el escenario correspondiente a ciertas aplicaciones con datos biológicos de interés actual, como la genómica y el microbioma, cada vez más importantes para la comprensión de ciertas enfermedades.

La reducción de dimensiones es una herramienta generalmente muy apropiada para abordar estas problemáticas, ya que puede ayudarnos a entender mejor los datos medidos. Consiste en obtener representaciones de los datos en un espacio de dimensión menor

que p, ya sea a través de una transformación adecuada o mediante la selección de variables relevantes. El objetivo principal es favorecer el análisis exploratorio y permitir un tratamiento estadístico más eficiente utilizando métodos clásicos.

Cuando el interés reside en describir o predecir una variable $Y \in \mathbb{R}$ en función de un conjunto de variables predictoras $\mathbf{X}$ en $\mathbb{R}^p$, reducir la dimensión de $\mathbf{X}$ puede simplificar el análisis siempre que no se comprometa la información relevante que $\mathbf{X}$ proporciona sobre Y. En este contexto, la *reducción suficiente de dimensiones* (SDR) es una metodología supervisada que busca proporcionar una reducción eficaz tanto en problemas de regresión como de clasificación, explotando el concepto de suficiencia estadística [Fisher, 1922]. La idea central es hallar una transformación $\mathbf{R}(\mathbf{X})$ de dimensión $q \leq p$, de manera tal que el estudio de $Y|\mathbf{R}(\mathbf{X})$ sea equivalente al de $Y|\mathbf{X}$ pero con la ventaja de estar formulado en un espacio de dimensión posiblemente mucho menor.

La metodología de SDR para problemas de aprendizaje supervisado fue introducida en [Li, 1991] y formalizada luego en términos de distribuciones condicionales en [Cook, 1998]. Las propuestas iniciales [Li, 1991; Cook and Weisberg, 1991; Li and Wang, 2007; Bura and Cook, 2001] se basaron en funciones de momentos de la distribución de $\mathbf{X}|Y$, enfoque que se conoce como *regresión inversa*. Bajo este mismo enfoque, en [Cook, 2007] se propuso una reducción basada en modelos que muestra de forma más clara la relación con el concepto de suficiencia estadística. Luego, diferentes métodos de reducción basados en modelos fueron propuestos dentro del enfoque de regresión inversa, tales como [Cook and Forzani, 2008, 2009] para el caso normal y [Bura and Forzani, 2015] para el caso de distribuciones de contorno elíptico.

Una característica en común de los métodos arriba mencionados es que obtienen SDR lineales de $\mathbf{X}$; esto es, de la forma $\mathbf{R}(\mathbf{X}) = \mathbf{\Gamma}^{\mathrm{T}}\mathbf{X}$, con $\mathbf{\Gamma} \in \mathbb{R}^{p\times q}$. La mayor limitación de los métodos lineales reside en que, en muchos problemas reales, la información relevante no puede suponerse que subyace en un subespacio afín de baja dimensión. Más aún, aunque sea posible encontrar un subespacio lineal que conserve la mayor parte de la información predictiva, en general esta dimensión es grande y, en consecuencia, se diluyen las potenciales ventajas de la reducción dimensional.

Un importante resultado de SDR no lineal basada en modelos fue presentado en [Bura et al., 2016], donde abordaron el caso general en que las distribuciones condicionales de $\mathbf{X}|Y$ pertenecen a una familia exponencial. En este estudio, hallaron de forma exhaustiva una SDR que no es lineal en $\mathbf{X}$ pero sí en el estadístico suficiente $\mathbf{T}(\mathbf{X})$ de la familia; es decir, es de la forma $\mathbf{R}(\mathbf{X}) = \mathbf{\Gamma}^{\mathrm{T}}\mathbf{T}(\mathbf{X})$. El uso de familias exponenciales también permitió obtener resultados de SDR para predictores mixtos [Bura et al., 2022].

Los intentos por obtener reducciones no lineales más generales son limitados. Una estrategia válida es transformar $\mathbf{X}$ a un espacio de mayor dimensión mediante un mapeo $\phi(\mathbf{X})$ y buscar allí direcciones de reducción. En este contexto, una herramienta que tuvo mucho éxito son los *espacios de Hilbert con núcleo reproductor* (RKHS) [Aronszajn, 1950], los cuales en general son de dimensión infinita y tienen la particularidad de que el mapeo ϕ queda determinado por una forma bilineal $k : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}$ denominada *núcleo reproductor*. Las propiedades de los núcleos reproductores permiten extender de forma natural algunos métodos clásicos de reducción lineal, obteniendo versiones no lineales de estos [Schölkopf et al., 1998; Mika et al., 1999; Wu, 2008].

Otra alternativa para la utilización de núcleos es a través de la definición de operadores de covarianza, los cuales permiten caracterizar los subespacios de reducción, conduciendo así a diferentes métodos de reducción basados en núcleos [Fukumizu et al., 2009; Kim and Pavlovic, 2011; Fukumizu and Leng, 2014]. Las principales ventajas de estos métodos radican en que realizan suposiciones débiles sobre la distribución de las variables y, además, en su proceso de estimación involucran únicamente matrices de Gram que se obtienen al evaluar el núcleo en la muestra.

Teniendo en cuenta lo descrito hasta aquí sobre el estado del arte en SDR, nuestro objetivo es aprovechar el uso de distribuciones condicionales dentro del enfoque de regresión inversa para obtener SDR no lineales en diferentes escenarios, preferentemente a partir de una familia de distribuciones lo suficientemente amplia y abarcativa para abordar problemas de alta complejidad. Una familia con tales características es la denominada *familia exponencial basada en núcleos* (KEF) [Canu and Smola, 2006; Fukumizu, 2009], la cual es una extensión infinito-dimensional de la familia exponencial.

1.1 Contribución de la tesis

Motivados por las estrategias basadas en regresión inversa de [Bura et al., 2016, 2022], en primer lugar estudiamos el problema de hallar una SDR de $\mathbf{X}$ cuando las distribuciones condicionales de $\mathbf{X}|Y$ pertenecen a una KEF. Bajo dicho modelo de regresión inversa, hallamos una expresión exhaustiva para una SDR de $\mathbf{X}$ cuando $Y|\mathbf{X}$ es un problema de clasificación. Además, proponemos un método de estimación eficiente basado en una conexión directa con el método de clasificación SVM [Boser et al., 1992].

Luego, estudiamos la teoría emergente de *reducción suficiente de dimensiones con información adicional* [Hung et al., 2015], en la cual se analiza el escenario predictivo donde, además de $\mathbf{X}$, existe otra variable $\mathbf{W}$ que contiene información sobre Y pero solo puede ser utilizada durante el proceso de estimación debido a diversas razones. En este contexto, aportamos una metodología general que aprovecha el potencial de los métodos basados en núcleos para manejar eficazmente una $\mathbf{W}$ de alta dimensionalidad.

1.1.1 Publicaciones y participaciones científicas

Los resultados principales sobre reducción suficiente de dimensiones para la familia exponencial basada en núcleos han sido publicados en el siguiente artículo científico:

> Ibañez, I., Forzani, L. and Tomassi, D. (2022). Generalized discriminant analysis via kernel exponential families. *Pattern Recognition*, 132:108933, ISSN 0031-3203, `https://doi.org/10.1016/j.patcog.2022.108933`

Además, resultados preliminares fueron comunicados en dos congresos de la especialidad:

> Supervised learning with kernel exponential families. *Workshop on Functional Inference and Machine Intelligence FIMI2020*, 17 al 19 de febrero de 2020, Sophia Antipolis, Francia.
>
> Supervised learning with infinite-dimensional kernel exponential families. *Primeras Jornadas de Inteligencia Artificial del Litoral*, 28 y 29 de noviembre de 2019, Santa Fe, Argentina.

Por su parte, los aportes correspondientes a reducción suficiente de dimensiones con información adicional conforman el siguiente artículo en proceso:

> Ibañez, I., Forzani, L. and Tomassi, D. (2023) Generalized kernel-based method for sufficient dimension reduction with additional information.

1.2 Organización de la tesis

Esta tesis está organizada de la siguiente manera:

- En los Capítulos 2 y 3 desarrollaremos el marco teórico. En el Capítulo 2 presentaremos un breve resumen de la teoría de núcleos reproductores y cómo definir a partir de ellos la *familia exponencial basada en núcleos*. En el Capítulo 3 introduciremos los conceptos básicos sobre *reducción suficiente de dimensiones* y repasaremos algunos métodos clásicos del estado del arte, los cuales están basados en transformaciones lineales o no lineales de las variables predictoras.

- En el Capítulo 4 identificaremos reducciones suficientes de dimensiones para la familia exponencial basada en núcleos. Luego, analizaremos formas alternativas de estimación; en particular, estableceremos una conexión directa con el conocido método de clasificación SVM [Boser et al., 1992], lo cual derivará en nuestro método propuesto denominado RKEF. Implementaremos y evaluaremos RKEF en simulaciones y datos reales a lo largo del Capítulo 5.

- En el Capítulo 6 estudiaremos el enfoque de *reducción suficiente de dimensiones con información adicional*, bajo el cual propondremos un método generalizado que permite combinar un par de métodos de reducción para lograr el propósito de incorporar la información extra. Mostraremos ejemplos al final del capítulo.

- Como cierre de la tesis, agruparemos algunos conceptos útiles en el Anexo A, mientras que las demostraciones de los resultados presentados durante el desarrollo del trabajo conformarán los Anexos B, C y D.

CAPÍTULO 2

Familia exponencial basada en núcleos

Muchas de las distribuciones que usamos para modelar el comportamiento de una variable aleatoria (v.a.) $\mathbf{X} \in \mathbb{R}^p$ se caracterizan por pertenecer a las denominadas *familias exponenciales* (EF). La importancia de estas familias radica en que son muy generales y existen formas eficientes ya estudiadas acerca de cómo extraer, de una muestra dada, información pertinente a los parámetros que caracterizan la familia. Dichos parámetros pueden ser representados por un vector $\boldsymbol{\theta} \in \mathbb{R}^m$.

Contar con un modelo estadístico tiene la ventaja de facilitar el análisis y la comprensión del comportamiento de las variables, así como las posibles relaciones entre ellas. Sin embargo, al adoptar tales modelos, generalmente se introducen condiciones que pueden resultar restrictivas, limitando el rango de aplicación. Por ese motivo, es constructivo proponer ideas cada vez más generales sin dejar de lado una cierta estructura predefinida. Dentro de este contexto, surge la posibilidad de extender el concepto de EF reemplazando de forma natural el uso de vectores $\boldsymbol{\theta} \in \mathbb{R}^m$ por funciones reales $f : \mathbb{R}^p \to \mathbb{R}$ en el rol de parámetro. Esta estrategia conduce a una formulación infinito-dimensional de la familia exponencial, la cual depende del espacio de funciones que se elija.

En particular, en [Canu and Smola, 2006; Fukumizu et al., 2009] proponen utilizar funciones pertenecientes a un espacio de Hilbert $\mathcal{H}_\mathcal{X}$, asociado de manera biunívoca a una función núcleo $k : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}$. Dichos espacios, denominados *espacios de Hilbert con núcleo reproductor* (RKHS) [Aronszajn, 1950], son una herramienta muy utilizada

dentro del análisis estadístico, debido esencialmente a que las propiedades de los núcleos reproductores han permitido formular diversas metodologías para extraer información predictiva. La extensión infinito-dimensional de EF obtenida se denomina *familia exponencial basada en núcleos* (KEF).

En este capítulo introduciremos el concepto de KEF, la cual constituye una clase muy amplia de distribuciones con la cual trabajaremos en esta tesis. Su definición y los resultados más importantes conformarán la Sección 2.3, pero antes repasaremos las características principales de los RKHS y haremos una breve reseña sobre las EF en las Secciones 2.1 y 2.2, respectivamente.

2.1 Espacios de Hilbert con núcleo reproductor (RKHS)

Comenzaremos introduciendo algunas notaciones. En un conjunto $\mathcal{X}$ no vacío se definen los espacios $L^p(\mathcal{X}, \mu)$ $(1 \leq p < \infty)$ de las funciones definidas en $\mathcal{X}$ tales que $\|f\|_p := \left(\int_{\mathcal{X}} |f|^p \, d\mu\right)^{1/p} < \infty$ (simplemente $L^p(\mathcal{X})$ si μ es la medida de Lebesgue), el espacio $L^\infty(\mathcal{X})$ de las funciones definidas en $\mathcal{X}$ tales que $\|f\|_\infty := \operatorname{ess\,sup} |f| < \infty$, donde $\operatorname{ess\,sup} |f| := \text{ínf}\{C \in \mathbb{R} : |f(x)| \leq C \text{ en casi todo punto } x \in \mathcal{X}\}$, y el espacio $C(\mathcal{X})$ de las funciones continuas definidas en $\mathcal{X}$. Además, para $\mathcal{X}$ Hausdorff localmente compacto[1], se define el espacio $C_0(\mathcal{X})$ de funciones continuas que se anulan en el infinito (esto es, para todo $\varepsilon > 0$ el conjunto $\{x \in \mathcal{X} : |f(x)| \geq \varepsilon\}$ es compacto), dotado con la norma uniforme $\|f\|_u := \sup_{x \in \mathcal{X}} |f(x)|$.

Para $\mathbf{u} = (u_1, \ldots, u_n) \in \mathbb{R}^n$ y $\mathbf{v} = (v_1, \ldots, v_n) \in \mathbb{R}^n$, el producto interno usual en $\mathbb{R}^n$ es $\langle \mathbf{u}, \mathbf{v} \rangle_{\mathbb{R}^n} := \sum_{i=1}^n u_i v_i$, mientras que la norma p, con $1 \leq p < \infty$, está dada por $\|\mathbf{u}\|_p := (|u_1|^p + \cdots + |u_n|^p)^{1/p}$; en particular, resulta $\|\mathbf{u}\|_2^2 = \langle \mathbf{u}, \mathbf{u} \rangle_{\mathbb{R}^n}$.

Por último, denotaremos span A el espacio generado por A o el espacio columna de A, según sea A una base o una matriz, respectivamente.

[1]Ejemplos de espacios de Hausdorff localmente compactos incluyen $\mathbb{R}^p$, sus subconjuntos abiertos o cerrados, los conjuntos discretos infinitos y las variedades topológicas.

La teoría general de núcleos fue desarrollada en [Aronszajn, 1943, 1950], aunque los núcleos reproductores han sido explorados desde principios del siglo XX. Iniciaremos definiendo los conceptos fundamentales de la teoría que son necesarias en esta tesis.

Definición 2.1. [Berlinet and Thomas-Agnan, 2004, Definición 1] Sea $\mathcal{H}_\mathcal{X}$ un espacio de Hilbert de funciones reales definidas en $\mathcal{X}$, con producto interno $\langle\cdot,\cdot\rangle_{\mathcal{H}_\mathcal{X}}$ y norma inducida[2] $\|\cdot\|_{\mathcal{H}_\mathcal{X}}$. Una función real $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ es un *núcleo reproductor* de $\mathcal{H}_\mathcal{X}$ si y solo si verifica las siguientes condiciones:

(I) Para todo $x \in \mathcal{X}$, $k(x,\cdot) \in \mathcal{H}_\mathcal{X}$.

(II) (*Propiedad reproductora*) Para todo $x \in \mathcal{X}$ y toda $f \in \mathcal{H}_\mathcal{X}$,

$$f(x) = \langle f, k(x,\cdot)\rangle_{\mathcal{H}_\mathcal{X}}. \tag{2.1}$$

En caso de admitir un núcleo reproductor, $\mathcal{H}_\mathcal{X}$ se denomina *espacio de Hilbert con núcleo reproductor* (RKHS).

La condición (I) de la Definición 2.1 implica que existe un mapeo $x \mapsto k(x,\cdot)$ que asocia a cada elemento de $\mathcal{X}$ una función en $\mathcal{H}_\mathcal{X}$. A $\phi(x) := k(x,\cdot)$ se lo denomina *mapeo característico*, mientras que $\mathcal{H}_\mathcal{X}$ es el *espacio característico*. A modo de ejemplo, veamos que $\mathbb{R}^n$ es un RKHS de dimensión finita.

Ejemplo 2.2. Sea $I_n = \{1,\dots,n\}$, con $n \in \mathbb{N}$. Podemos identificar un vector $\mathbf{x} = (x_1,\dots,x_n) \in \mathbb{R}^n$ con la función $f_\mathbf{x} : I_n \to \mathbb{R}$ definida por $f_\mathbf{x}(i) = x_i$. El par $(\mathcal{H}, \langle\cdot,\cdot\rangle_\mathcal{H})$, donde

$$\mathcal{H} = \{f_\mathbf{x} : \mathbf{x} \in \mathbb{R}^n\} \qquad \text{y} \qquad \langle f_\mathbf{x}, f_\mathbf{y}\rangle_\mathcal{H} = \langle \mathbf{x}, \mathbf{y}\rangle_{\mathbb{R}^n},$$

es un RKHS con núcleo reproductor $k : I_n \times I_n \to \mathbb{R}$ definido por

$$k(i,j) = \delta_{ij} := \begin{cases} 1 & \text{si } i = j, \\ 0 & \text{si } i \neq j. \end{cases}$$

[2]Esto es, $\|f\|_{\mathcal{H}_\mathcal{X}} := \langle f, f\rangle_{\mathcal{H}_\mathcal{X}}^{1/2}$ para toda $f \in \mathcal{H}_\mathcal{X}$.

Veamos que se verifican las condiciones (I) y (II) de la Definición 2.1. Si $\mathbf{e}_{i_0}$ es el vector canónico de $\mathbb{R}^n$ con valor 1 en la i_0-ésima coordenada, para todo $i_0 \in I_n$ y toda $f_{\mathbf{x}} \in \mathcal{H}$:

(I) $k(i_0, \cdot) = f_{\mathbf{e}_{i_0}} \in \mathcal{H}$.

(II) $f_{\mathbf{x}}(i_0) = \langle \mathbf{x}, \mathbf{e}_{i_0} \rangle_{\mathbb{R}^n} = \langle f_{\mathbf{x}}, k(i_0, \cdot) \rangle_{\mathcal{H}}$. □

Una propiedad muy importante de los RKHS se obtiene combinando (I) y (II): para todo $x_1, x_2 \in \mathcal{X}$ se verifica

$$k(x_1, x_2) = k(x_2, x_1) = \langle k(x_1, \cdot), k(x_2, \cdot) \rangle_{\mathcal{H}_\mathcal{X}}. \tag{2.2}$$

La expresión (2.2) no solo nos informa que todo núcleo reproductor k es simétrico, sino que además nos dice que k contiene información sobre el producto interno en $\mathcal{H}_\mathcal{X}$. En efecto, si $f = \sum_{i=1}^n a_i k(x_i, \cdot)$ y $g(x) = \sum_{j=1}^m b_j k(\tilde{x}_j, \cdot)$, con $x_i, \tilde{x}_j \in \mathcal{X}$ y $a_i, b_j \in \mathbb{R}$, entonces tanto f como g pertenecen a $\mathcal{H}_\mathcal{X}$ y, por propiedades de producto interno y (2.2), se obtiene

$$\langle f, g \rangle_{\mathcal{H}_\mathcal{X}} = \sum_{i=1}^n \sum_{j=1}^m a_i b_j k(x_i, \tilde{x}_j) \qquad \text{y} \qquad \|f\|^2_{\mathcal{H}_\mathcal{X}} = \sum_{i=1}^n \sum_{j=1}^n a_i a_j k(x_i, x_j). \tag{2.3}$$

Ahora que los RKHS están definidos, analicemos las siguientes dos preguntas:

(a) Dado un RKHS $\mathcal{H}_\mathcal{X}$, ¿su núcleo reproductor k es único?

(b) ¿Qué condiciones debe tener una aplicación $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ para ser núcleo reproductor de algún RKHS? ¿Dicho RKHS es único?

La respuesta a (a) es afirmativa, por consecuencia directa de la propiedad reproductora (2.1). En efecto, si k_1 y k_2 son núcleos reproductores de $\mathcal{H}_\mathcal{X}$, para toda $f \in \mathcal{H}_\mathcal{X}$ y todo $x \in \mathcal{X}$, se tiene que

$$\langle f, k_1(x, \cdot) - k_2(x, \cdot) \rangle_{\mathcal{H}_\mathcal{X}} = \langle f, k_1(x, \cdot) \rangle_{\mathcal{H}_\mathcal{X}} - \langle f, k_2(x, \cdot) \rangle_{\mathcal{H}_\mathcal{X}} = f(x) - f(x) = 0.$$

En particular, tomando $f = k_1(x, \cdot) - k_2(x, \cdot) \in \mathcal{H}_\mathcal{X}$ se deduce $\|k_1(x, \cdot) - k_2(x, \cdot)\|_{\mathcal{H}_\mathcal{X}} = 0$. Luego, debe ser $k_1(x, \cdot) = k_2(x, \cdot)$ para todo $x \in \mathcal{X}$.

En cuanto a la pregunta (b), la clave está en la segunda expresión de (2.3): dado que, por definición de norma, $\|f\|_{\mathcal{H}_\mathcal{X}} \geq 0$, se deduce que $\sum_{i=1}^n \sum_{j=1}^n a_i a_j k(x_i, x_j) \geq 0$ para todo núcleo reproductor k y para cualquier elección de $\{x_i\}_{i=1}^n \subset \mathcal{X}$ y $(a_1, \ldots, a_n) \in \mathbb{R}^n$. Esto conduce a la siguiente definición.

Definición 2.3. [Aronszajn, 1950] Una función $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ es *definida positiva en el sentido de E.H. Moore* si, para todo $\{x_i\}_{i=1}^n \subset \mathcal{X}$ y todo $(a_1, \ldots, a_n) \in \mathbb{R}^n$, se verifica

$$\sum_{i=1}^n \sum_{j=1}^n a_i a_j k(x_i, x_j) \geq 0.$$

Podemos afirmar que todo núcleo reproductor k es una función definida positiva en el sentido de E.H. Moore. Sumado al hecho de que los núcleos reproductores son simétricos, diremos de forma más abreviada que un núcleo reproductor k es una función *simétrica definida positiva* (s.d.p.). El recíproco, conocido como Teorema de Moore-Aronszajn, es fundamental en la teoría de los RKHS, ya que proporciona una forma de elegir funciones $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ válidas como núcleos reproductores de *algún* espacio de Hilbert $\mathcal{H}_\mathcal{X}$. Además, asegura que el RKHS que induce un núcleo reproductor es único.

Teorema 2.4. (Teorema de Moore-Aronszajn) [Berlinet and Thomas-Agnan, 2004, Teorema 3] *Sea $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ una función s.d.p. Existe un único RKHS $\mathcal{H}_\mathcal{X}$ con k como núcleo reproductor. Además, el subespacio*

$$\mathcal{H}_0 := \operatorname{span}\{k(x, \cdot) : x \in \mathcal{X}\}$$

es denso en $\mathcal{H}_\mathcal{X}$ y $\mathcal{H}_\mathcal{X}$ es el conjunto de funciones en $\mathcal{X}$ que son límites puntuales de sucesiones de Cauchy[3] *en $\mathcal{H}_0$ con el producto interno*

$$\langle f, g \rangle_{\mathcal{H}_0} := \sum_{i=1}^n \sum_{j=1}^m a_i b_j k(x_i, \tilde{x}_j),$$

donde $f = \sum_{i=1}^n a_i k(x_i, \cdot)$ y $g = \sum_{j=1}^m b_j k(\tilde{x}_j, \cdot)$.

[3] $\{f_n\} \subset \mathcal{H}_0$ es de Cauchy si, dado $\varepsilon > 0$, existe $N \in \mathbb{N}$ tal que $\|f_n - f_m\|_{\mathcal{H}_0} < \varepsilon$ para todo $n, m \geq N$.

Núcleo	Expresión de $k(\mathbf{x}_1, \mathbf{x}_2)$	Parámetro/s
Núcleo lineal	$\langle \mathbf{x}_1, \mathbf{x}_2 \rangle_{\mathbb{R}^n}$	No tiene
Núcleo polinomial	$(c + \gamma \langle \mathbf{x}_1, \mathbf{x}_2 \rangle_{\mathbb{R}^n})^d$	$c \in \mathbb{R}$, $\gamma \in \mathbb{R}$, $d \in \mathbb{N}$
Núcleo Gaussiano	$\exp\left\{-\frac{\|\mathbf{x}_1 - \mathbf{x}_2\|_2^2}{\sigma^2}\right\}$	$\sigma \in \mathbb{R}^+$
Núcleo Laplaciano	$\exp\left\{-\frac{\|\mathbf{x}_1 - \mathbf{x}_2\|_1}{\sigma^2}\right\}$	$\sigma \in \mathbb{R}^+$
Núcleo multicuadrático inverso	$\left(1 + \frac{\|\mathbf{x}_1 - \mathbf{x}_2\|_2^2}{c^2}\right)^{-\beta}$	$\beta \in \mathbb{R}^+$, $c \in \mathbb{R}^+$

TABLA 2.1. Ejemplos de núcleos reproductores en $\mathbb{R}^n$.

En la Tabla 2.1 se presentan los núcleos reproductores más utilizados para $\mathcal{X} = \mathbb{R}^n$. Dichos núcleos pueden depender de uno o más parámetros en su definición y, en general, están asociados a un RKHS de alta dimensión, posiblemente infinita. El núcleo Gaussiano es el más famoso y el más utilizado en aplicaciones, debido a que las propiedades que posee permiten que verifique en general las suposiciones bajo las cuales se aseguran los resultados teóricos de los diferentes métodos basados en núcleos. A continuación haremos un breve resumen sobre sus principales características.

2.1.1 Núcleo Gaussiano

En $\mathcal{X} = \mathbb{R}^n$, el núcleo Gaussiano

$$k(\mathbf{x}_1, \mathbf{x}_2) = \exp\left\{-\frac{\|\mathbf{x}_1 - \mathbf{x}_2\|^2}{\sigma^2}\right\} \tag{2.4}$$

genera un RKHS de dimensión infinita definido por

$$\mathcal{H}_{\mathcal{X},\sigma} := \left\{ f \in L^2(\mathbb{R}^n) \cap C(\mathbb{R}^n) : \int |f^\wedge(\boldsymbol{\omega})|^2 \exp\{\sigma^2 \|\boldsymbol{\omega}\|_2^2/4\}\, d\boldsymbol{\omega} < \infty \right\},$$

donde $f^\wedge(\boldsymbol{\omega}) := (2\pi)^{-p/2} \int_{\mathbb{R}^p} f(\mathbf{x}) \exp\{-i\langle \boldsymbol{\omega}, \mathbf{x} \rangle_{\mathbb{R}^p}\}\, d\mathbf{x}$ es la transformada de Fourier de f. Para un estudio detallado de dicho RKHS, ver [Steinwart et al., 2006; Steinwart and Christmann, 2008].

Las características principales de un núcleo reproductor tienen que ver con las nociones de núcleo *universal* y/o *característico* [ver Sriperumbudur et al., 2011]. La universalidad

de un núcleo está relacionada a qué tan abarcativo es el RKHS que genera. Por su parte, los núcleos característicos permiten distinguir medidas de probabilidad embebidas dentro de un RKHS. En particular, el núcleo Gaussiano (2.4):

- Es *característico*. Si $M_+^1(\mathcal{X})$ es el espacio de las medidas de probabilidad sobre $\mathcal{X}$, entonces el mapeo $\mu : M_+^1(\mathcal{X}) \to \mathcal{H}_{\mathcal{X},\sigma}$ definido por $\mu(\mathbb{P}) = \mathbb{E}_{X\sim\mathbb{P}}[k(X,\cdot)]$ es inyectivo. Esto significa que μ caracteriza las medidas de probabilidad en $\mathcal{H}_{\mathcal{X}}$.
- Es *c universal*. El espacio $\mathcal{H}_{\mathcal{X},\sigma}$ es denso en $C(\mathcal{X})$ respecto a la norma uniforme; es decir, para toda $g \in C(\mathcal{X})$ y $\varepsilon > 0$, existe $f \in \mathcal{H}_{\mathcal{X},\sigma}$ tal que $\|f - g\|_u < \varepsilon$.
- Es c_0 *universal*. Es acotado, con $k(\mathbf{x},\cdot) \in C_0(\mathcal{X})$ para todo $\mathbf{x} \in \mathcal{X}$, y además $\mathcal{H}_{\mathcal{X},\sigma}$ es denso en $C_0(\mathcal{X})$ respecto a la norma uniforme.

2.2 Familia exponencial (EF)

De ahora en adelante, consideraremos $\mathbf{X} \in \mathcal{X} \subset \mathbb{R}^p$ una v.a. en un espacio de probabilidad $\mathscr{P}$. Además, con $\mathcal{D}_x$ nos referiremos a un conjunto de n datos de una v.a. X; en particular, $\mathcal{D}_{\mathbf{x}} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$.

Dada una matriz $\mathbf{A} \in \mathbb{R}^{n\times m}$, $\operatorname{vec}\mathbf{A} := (\mathbf{a}_{\cdot 1}^{\mathrm{T}}, \mathbf{a}_{\cdot 2}^{\mathrm{T}}, \ldots, \mathbf{a}_{\cdot m}^{\mathrm{T}})^{\mathrm{T}} \in \mathbb{R}^{nm}$ es su vectorización, donde $\mathbf{a}_{\cdot j}$ es la j-ésima columna de $\mathbf{A}$. Si $\mathbf{A} \in \mathbb{R}^{n\times n}$ es simétrica, $\operatorname{vech}\mathbf{A} \in \mathbb{R}^{n(n+1)/2}$ es la vectorización de su parte triangular superior. Por último, con $\mathbf{1}(\cdot)$ nos referiremos a la función indicatriz, que asume el valor 1 si el argumento es cierto y 0 en caso contrario.

Iniciaremos repasando el concepto de EF, que extenderemos en la próxima sección a un contexto de dimensión infinita.

Definición 2.5. Una v.a. $\mathbf{X} \in \mathbb{R}^p$ está distribuida en una *familia exponencial* (EF) *a m parámetros* si su función de densidad pertenece al conjunto

$$\mathcal{P}_{\text{fin}} := \left\{ p(\mathbf{x}|\boldsymbol{\theta}) = \frac{q_0(\mathbf{x})}{Z(\boldsymbol{\theta})} \exp\langle \boldsymbol{\theta}, \mathbf{T}(\mathbf{x})\rangle_{\mathbb{R}^m} : \boldsymbol{\theta} \in \boldsymbol{\Theta} \right\}, \tag{2.5}$$

donde $q_0 : \mathbb{R}^p \to \mathbb{R}^+$, $Z(\boldsymbol{\theta}) := \int_{\mathcal{X}} q_0(\mathbf{x}) \exp\langle \boldsymbol{\theta}, \mathbf{T}(\mathbf{x})\rangle_{\mathbb{R}^m} d\mathbf{x}$ y $\boldsymbol{\Theta} := \{\boldsymbol{\theta} \in \mathbb{R}^m : Z(\boldsymbol{\theta}) < \infty\}$.

En la Definición 2.5 se identifican el *parámetro natural* $\boldsymbol{\theta}$ y el *estadístico suficiente* $\mathbf{T}(\mathbf{x})$, ambos de la misma dimensión. Muchas distribuciones conocidas pertenecen a una EF, tanto de variables continuas como discretas (en cuyo caso la definición de $Z(\boldsymbol{\theta})$ es en término de sumatoria). Algunos ejemplos en el caso univariado $p = 1$ son las distribuciones Normal, Bernoulli y Poisson, mientras que en el caso multivariado $p > 1$ se destacan las distribuciones Normal Multivariada y Multinomial. A continuación, identificaremos $\boldsymbol{\theta}$ y $\mathbf{T}(\mathbf{x})$ en las distribuciones multivariadas mencionadas.

Ejemplo 2.6. Sea $\mathbf{X} \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Delta})$, con $\boldsymbol{\mu} \in \mathbb{R}^p$ y $\boldsymbol{\Delta} \in \mathbb{R}^{p\times p}$ s.d.p. (se denota $\boldsymbol{\Delta} > 0$). La función de densidad de $\mathbf{X}$ es

$$\begin{aligned} p(\mathbf{x}) &= \frac{1}{(2\pi)^{n/2}|\boldsymbol{\Delta}|^{1/2}} \exp\left\{-\frac{1}{2}\langle \mathbf{x}-\boldsymbol{\mu}, \boldsymbol{\Delta}^{-1}(\mathbf{x}-\boldsymbol{\mu})\rangle_{\mathbb{R}^p}\right\} \\ &= \exp\left\{\langle \boldsymbol{\Delta}^{-1}\boldsymbol{\mu}, \mathbf{x}\rangle_{\mathbb{R}^p} - \frac{1}{2}\langle \mathbf{x}, \boldsymbol{\Delta}^{-1}\mathbf{x}\rangle_{\mathbb{R}^p} - \tilde{Z}(\boldsymbol{\mu}, \boldsymbol{\Delta})\right\} \\ &= \exp\left\{\langle \boldsymbol{\Delta}^{-1}\boldsymbol{\mu}, \mathbf{x}\rangle_{\mathbb{R}^p} + \langle \operatorname{vec}\boldsymbol{\Delta}^{-1}, -\frac{1}{2}\operatorname{vec}(\mathbf{x}\mathbf{x}^{\mathrm{T}})\rangle_{\mathbb{R}^{p^2}} - \tilde{Z}(\boldsymbol{\mu}, \boldsymbol{\Delta})\right\} \\ &= \exp\left\{\langle \boldsymbol{\Delta}^{-1}\boldsymbol{\mu}, \mathbf{x}\rangle_{\mathbb{R}^p} + \langle \operatorname{vech}\boldsymbol{\Delta}^{-1}, -\frac{1}{2}\mathbf{D}_p^{\mathrm{T}}\mathbf{D}_p\operatorname{vech}(\mathbf{x}\mathbf{x}^{\mathrm{T}})\rangle_{\mathbb{R}^{p(p+1)/2}} - \tilde{Z}(\boldsymbol{\mu}, \boldsymbol{\Delta})\right\}, \end{aligned}$$

donde

$$\tilde{Z}(\boldsymbol{\mu}, \boldsymbol{\Delta}) = \frac{1}{2}\left[\langle \boldsymbol{\mu}, \boldsymbol{\Delta}^{-1}\boldsymbol{\mu}\rangle_{\mathbb{R}^p} + \log|\boldsymbol{\Delta}| + p\log(2\pi)\right]$$

y $\mathbf{D}_p$ es una matriz tal que $\mathbf{D}_p \operatorname{vech}\mathbf{A} = \operatorname{vec}\mathbf{A}$ para toda $\mathbf{A} \in \mathbb{R}^{p\times p}$ simétrica. Entonces $\mathbf{X}$ se distribuye en una EF a $p + p(p+1)/2$ parámetros, con estadístico suficiente

$$\mathbf{T}(\mathbf{x}) = \left(\mathbf{x}, -\frac{1}{2}\mathbf{D}_p^{\mathrm{T}}\mathbf{D}_p \operatorname{vech}(\mathbf{x}\mathbf{x}^{\mathrm{T}})\right) \tag{2.6}$$

y parámetro natural

$$\boldsymbol{\theta} = (\boldsymbol{\Delta}^{-1}\boldsymbol{\mu}, \operatorname{vech}\boldsymbol{\Delta}^{-1}). \tag{2.7}$$

□

Ejemplo 2.7. Sea $\mathbf{X} = (X_1, \ldots, X_p)$, con $X_i \in \{1, \ldots, r\}$ y $\sum_{i=1}^p X_i = r$, una v.a. que describe la ocurrencia de p eventos excluyentes luego de r ensayos; esto es, X_i es la cantidad de veces que ocurre el i-ésimo evento. Entonces $\mathbf{X}$ tiene distribución Multinomial

y su función de probabilidad es

$$\mathbb{P}\{\mathbf{X}=\mathbf{x}\} = \frac{r!}{x_1!x_2!\dots x_p!} p_1^{x_1} p_2^{x_2} \dots p_r^{x_r}\ \mathbf{1}\left(\|\mathbf{x}\|_1 = r\right), \tag{2.8}$$

donde $0 < p_k < 1$ $(k = 1, \dots, p)$ son las probabilidades de ocurrencia de los eventos, de manera tal que $\sum_{i=1}^{p} p_i = 1$. La expresión (2.8) puede reescribirse como

$$\begin{aligned}
\mathbb{P}\{\mathbf{X}=\mathbf{x}\} &= \exp\left\{\sum_{i=1}^{r} x_i \log p_i\right\} \frac{r!}{x_1!\dots x_p!}\ \mathbf{1}\left(\|\mathbf{x}\|_1 = r\right) \\
&= \exp\left\{\sum_{i=1}^{r-1} x_i \log p_i + \left(r - \sum_{i=1}^{r-1} x_i\right) \log\left(1 - \sum_{i=1}^{r-1} p_i\right)\right\} \frac{r!}{x_1!\dots x_p!}\ \mathbf{1}\left(\|\mathbf{x}\|_1 = r\right) \\
&= \exp\left\{\sum_{i=1}^{r-1} x_i \log \frac{p_i}{1 - \sum_{j=1}^{r-1} p_i} + r \log\left(1 - \sum_{i=1}^{r-1} p_i\right)\right\} \frac{r!}{x_1!\dots x_p!}\ \mathbf{1}\left(\|\mathbf{x}\|_1 = r\right).
\end{aligned}$$

Por lo tanto, $\mathbf{X}$ se distribuye en una EF a $r-1$ parámetros, con estadístico suficiente $\mathbf{T}(\mathbf{x}) = (x_1, \dots, x_{r-1})$ y parámetro natural

$$\boldsymbol{\theta} = \left(\log \frac{p_1}{1 - \sum_{i=1}^{r-1} p_j}, \dots, \log \frac{p_{r-1}}{1 - \sum_{i=1}^{r-1} p_j}\right).$$

□

La riqueza de las EF va más allá de los casos mencionados anteriormente. Una aplicación importante es el de la mezcla de variables categóricas y continuas, que suelen presentarse en conjuntos de datos de diversos campos, como la genómica y las finanzas. Suponiendo $\mathbf{X} = (\tilde{\mathbf{X}}_1, \tilde{\mathbf{X}}_2)$, con $\tilde{\mathbf{X}}_1 \in \mathbb{R}^{p_1}$ un vector de variables binarias y $\tilde{\mathbf{X}}_2 \in \mathbb{R}^{p_2}$ un vector de variables continuas, bajo ciertos modelos para las distribuciones de $\tilde{\mathbf{X}}_1$ y $\tilde{\mathbf{X}}_2|\tilde{\mathbf{X}}_1$ resulta que $\mathbf{X}$ está distribuida en una EF [Bura et al., 2022].

2.3 Familia exponencial basada en núcleos (KEF)

En lo que sigue, para dos funciones de densidad p y q en $\mathcal{X}$, $D_{\mathrm{KL}}(p, q)$ y $h(p, q)$ denotan la divergencia de Kullback-Leibler y la distancia de Hellinger, respectivamente, a saber:

$$D_{\mathrm{KL}}(p, q) := \int_{\mathcal{X}} p(\mathbf{x}) \log \frac{p(\mathbf{x})}{q(\mathbf{x})}\, d\mathbf{x}, \tag{2.9}$$

$$h(p, q) := \frac{1}{\sqrt{2}} \|\sqrt{p} - \sqrt{q}\|_2.$$

A continuación, definiremos la extensión infinito-dimensional de la familia exponencial. En dicha extensión, el rol de parámetro natural lo asume una función $f : \mathbb{R}^p \to \mathbb{R}$ perteneciente a un RKHS posiblemente de dimensión infinita.

Definición 2.8. [Canu and Smola, 2006; Fukumizu, 2009] Sea $\mathcal{H}_\mathcal{X}$ un RKHS con núcleo reproductor $k : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}$. Una v.a. $\mathbf{X} \in \mathbb{R}^p$ está distribuida en la *familia exponencial basada en núcleos* (KEF) generada por $\mathcal{H}_\mathcal{X}$ si su función de densidad pertenece al conjunto

$$\mathcal{P} := \left\{ p_f(\mathbf{x}) = \frac{q_0(\mathbf{x})}{Z(f)} \exp\langle f, k(\mathbf{x}, \cdot)\rangle_{\mathcal{H}_\mathcal{X}} : f \in \mathcal{F} \right\}, \tag{2.10}$$

donde $q_0 : \mathbb{R}^p \to \mathbb{R}^+$, $Z(f) := \int_\mathcal{X} q_0(\mathbf{x}) \exp\langle f, k(\mathbf{x}, \cdot)\rangle_{\mathcal{H}_\mathcal{X}} \, d\mathbf{x}$ y $\mathcal{F} := \{f \in \mathcal{H}_\mathcal{X} : Z(f) < \infty\}$.

Comparando (2.10) con (2.5), podemos decir que f y $k(\mathbf{x}, \cdot)$ asumen el rol de parámetro natural y estadístico suficiente, respectivamente. Además, se dice que $\mathcal{P}$ es una extensión infinito-dimensional de $\mathcal{P}_{\text{fin}}$, ya que toda EF se puede escribir en forma de $\mathcal{P}$ mediante un RKHS de dimensión finita, tal como veremos en el siguiente ejemplo.

Ejemplo 2.9. Sea $\mathbf{X} \in \mathbb{R}^p$ una v.a. distribuida en $\mathcal{P}_{\text{fin}}$, con parámetro natural $\boldsymbol{\theta} \in \mathbb{R}^m$. Para escribir $\mathcal{P}_{\text{fin}}$ como $\mathcal{P}$, basta considerar el espacio de funciones reales en $\mathbb{R}^p$ dado por

$$\mathcal{H}_\mathcal{X} = \text{span}\left\{T_1(\mathbf{x}), \cdots, T_m(\mathbf{x})\right\}, \tag{2.11}$$

dotado con el producto interno $\langle f, g\rangle_{\mathcal{H}_\mathcal{X}} = \langle \boldsymbol{\alpha}, \boldsymbol{\beta}\rangle_{\mathbb{R}^m}$, para $f(\mathbf{x}) = \langle \boldsymbol{\alpha}, \mathbf{T}(\mathbf{x})\rangle_{\mathbb{R}^m}$ y $g(\mathbf{x}) = \langle \boldsymbol{\beta}, \mathbf{T}(\mathbf{x})\rangle_{\mathbb{R}^m}$. Entonces $\mathcal{H}_\mathcal{X}$ es un RKHS con núcleo reproductor

$$k(\mathbf{x}_1, \mathbf{x}_2) = \langle \mathbf{T}(\mathbf{x}_1), \mathbf{T}(\mathbf{x}_2)\rangle_{\mathbb{R}^m}. \tag{2.12}$$

En efecto, veamos que se cumplen las condiciones de la Definición 2.1:

(I) Para todo $\mathbf{x}_0 \in \mathcal{X}$, $k(\mathbf{x}, \cdot) = \langle \mathbf{T}(\mathbf{x}), \mathbf{T}(\cdot)\rangle_{\mathbb{R}^m}$. Por lo tanto, de (2.11) resulta $k(\mathbf{x}, \cdot) \in \mathcal{H}_\mathcal{X}$. Observar que en esta expresión queda explícita la correspondencia entre $k(\mathbf{x}, \cdot)$ y $\mathbf{T}(\mathbf{x})$ para el caso $\mathcal{P}_{\text{fin}}$.

(II) (Propiedad reproductora) Para todo $\mathbf{x} \in \mathcal{X}$ y toda $f = \langle \boldsymbol{\alpha}, \mathbf{T}(\cdot)\rangle_{\mathbb{R}^m} \in \mathcal{H}_{\mathcal{X}}$,

$$f(\mathbf{x}) = \langle \boldsymbol{\alpha}, \mathbf{T}(\mathbf{x})\rangle_{\mathbb{R}^m} = \langle f, k(\mathbf{x}, \cdot)\rangle_{\mathcal{H}_{\mathcal{X}}}.$$

Finalmente, comparando (2.5) y (2.10), podemos decir que $\mathbf{X}$ está distribuida en la familia $\mathcal{P}$ generada por el RKHS $\mathcal{H}_{\mathcal{X}}$ de dimensión finita, y su función de densidad es $p_f(\mathbf{x})$ con parámetro funcional $f = \langle \boldsymbol{\theta}, \mathbf{T}(\cdot)\rangle_{\mathbb{R}^m}$. □

Como caso particular y continuando el Ejemplo 2.6, veamos como expresar la distribución Normal multivariada como una KEF generada por un RKHS de dimensión finita.

Ejemplo 2.10. Sea $\mathbf{X} \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Delta})$, con $\boldsymbol{\mu} \in \mathbb{R}^p$ y $\boldsymbol{\Delta} \in \mathbb{R}^{p\times p}$ s.d.p. De (2.6), (2.11) y (2.12), se concluye que $\mathbf{X}$ está distribuida en la familia $\mathcal{P}$ correspondiente al RKHS $\mathcal{H}_{\mathcal{X}}$ de dimensión $p + p(p+1)/2$ generado por el núcleo reproductor

$$k(\mathbf{x}_1, \mathbf{x}_2) = \langle \mathbf{x}_1, \mathbf{x}_2\rangle_{\mathbb{R}^p} + \frac{1}{4}\langle \mathbf{D}_p^{\mathrm{T}}\mathbf{D}_p \mathrm{vech}\,(\mathbf{x}_1\mathbf{x}_1^{\mathrm{T}}), \mathbf{D}_p^{\mathrm{T}}\mathbf{D}_p \mathrm{vech}\,(\mathbf{x}_2\mathbf{x}_2^{\mathrm{T}})\rangle_{\mathbb{R}^{p(p+1)/2}}.$$

La función de densidad de $\mathbf{X}$ tiene parámetro funcional $f = \langle \boldsymbol{\theta}, \mathbf{T}(\cdot)\rangle_{\mathbb{R}^m} \in \mathcal{H}_{\mathcal{X}}$, donde $\mathbf{T}$ y $\boldsymbol{\theta}$ están dados por (2.6) y (2.7), respectivamente. □

A pesar de esta conexión entre EF y KEF, es importante remarcar que trabajar con un RKHS de dimensión infinita tiene la ventaja de que, bajo leves condiciones del núcleo reproductor k, se construye una familia que permite aproximar una amplia clases de distribuciones tanto como deseemos, en virtud del siguiente resultado.

Proposición 2.11. [Sriperumbudur et al., 2017, Proposición 1] *Sean $\mathcal{X} \subset \mathbb{R}^p$ un conjunto localmente compacto y*

$$\mathcal{P}_0 := \left\{ \pi_f(\mathbf{x}) = \frac{q_0(\mathbf{x})}{Z(f)} \exp\langle f, k(\mathbf{x}, \cdot)\rangle_{\mathcal{H}_{\mathcal{X}}} : f \in C_0(\mathcal{X}) \right\}. \tag{2.13}$$

Si $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ es c_0 universal, entonces $\mathcal{P}$ es densa en $\mathcal{P}_0$ respecto a la divergencia de Kullback-Leibler, a la distancia de Hellinger y a la norma L^1. Además, si para algún $1 < r \leq \infty$ se cumple que $q_0 \in L^1(\mathcal{X}) \cap L^r(\mathcal{X})$, $\mathcal{P}$ es densa en $\mathcal{P}_0$ respecto a la norma L^r.

La importancia de la Proposición 2.11 es que la familia $\mathcal{P}_0$ dada por (2.13) contiene una amplia clase de densidades continuas [Sriperumbudur et al., 2017, Corolario 2]. Más aún, si $\mathcal{X}$ es un conjunto cerrado y acotado de $\mathbb{R}^p$, y $q_0(\mathbf{x})$ es una distribución uniforme en $\mathcal{X}$, cualquier densidad continua en $\mathcal{X}$ puede ser aproximada tanto como se quiera mediante densidades en $\mathcal{P}$.

2.3.1 Estimación de densidades en KEF

En esta sección repasaremos un método de estimación de densidades en KEF, propuesto en [Sriperumbudur et al., 2017]. Asumamos que la función de densidad $p(\mathbf{x})$ pertenece a una familia $\mathcal{P}$ dada por (2.10); es decir, $p(\mathbf{x}) = p_{f_0}(\mathbf{x})$ para algún $f_0 \in \mathcal{H}_\mathcal{X}$. No obstante, en [Sriperumbudur et al., 2017] probaron que el estimador mediante densidades de $\mathcal{P}$ es efectivo también para el caso $p \notin \mathcal{P}$, lo cual es más razonable en la práctica.

La técnica de *estimación por máxima verosimilitud* (MLE), consistente en minimizar la divergencia de Kullback-Leibler (2.9), no es eficiente para estimar densidades en $\mathcal{P}$ y $\mathcal{P}_{\text{fin}}$. Esto es debido a las dificultades en el manejo de la función de partición Z, que en el caso de $\mathcal{P}$ está dada por

$$Z(f) := \int_\mathcal{X} q_0(\mathbf{x}) \exp\langle f, k(\mathbf{x}, \cdot)\rangle_{\mathcal{H}_\mathcal{X}} \, d\mathbf{x}.$$

Por ese motivo, en [Sriperumbudur et al., 2017] proponen aplicar la técnica de *score matching* [Hyvärinen, 2005], la cual consiste en minimizar la divergencia de Fisher definida por

$$J(p, p_f) := \frac{1}{2} \int_\mathcal{X} p(\mathbf{x}) \|\nabla \log p(\mathbf{x}) - \nabla \log p_f(\mathbf{x})\|_2^2 \, d\mathbf{x}.$$

Para ello, en primer lugar proponen regularizar el problema definiendo

$$J_\lambda(f) := J(p, p_f) + \frac{\lambda}{2} \|f\|_{\mathcal{H}_\mathcal{X}}^2, \quad \lambda > 0. \tag{2.14}$$

y enumeran una lista de suposiciones necesarias para el análisis [Sriperumbudur et al., 2017, suposiciones (A)-(D)], entre las que destacan:

- k dos veces continuamente diferenciable en $\mathcal{X} \times \mathcal{X}$. Resulta $f \in \mathcal{H}_\mathcal{X}$ dos veces continuamente diferenciable [Steinwart and Christmann, 2008, Corolario 4.36].

- (ε integrabilidad) Para algún $\varepsilon \geq 1$ y para todo $i = 1, \ldots, p$, se verifica

$$\frac{\partial}{\partial x_i}\frac{\partial}{\partial x_{i+p}}k(\mathbf{x}, \mathbf{x}) \in L^{\varepsilon}(\mathcal{X}, p),$$

  donde se usa $\frac{\partial}{\partial x_i}$ y $\frac{\partial}{\partial x_{j+p}}$ para indicar que se deriva respecto de la i-ésima componente de la 1° variable y de la j-ésima componente de la 2° variable, respectivamente. Además, si $\varepsilon = 1$, la condición asegura $J(p, p_f) < \infty$ para todo $p_f \in \mathcal{P}$.

Ejemplos de núcleos que satisfacen las condiciones dadas son el núcleo Gaussiano y el núcleo multicuadrático inverso, ambos definidos en la Tabla 2.1. En [Sriperumbudur et al., 2017, Teorema 4] construyen un estimador $\hat{f}_\lambda$ que depende de k y q_0, pero que es difícil de calcular en la práctica. Afortunadamente, proveen luego una expresión alternativa para $\hat{f}_\lambda$ que involucra un sistema de ecuaciones lineales.

Teorema 2.12. [Sriperumbudur et al., 2017, Teorema 5] *Sea $\mathcal{D}_{\mathbf{x}}$ un conjunto de datos de una v.a. $\mathbf{X} \in \mathbb{R}^p$ con función de densidad $p(\mathbf{x})$ en la familia $\mathcal{P}$ definida por (2.10), la cual está generada por un núcleo reproductor $k : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}$. Dado $\lambda > 0$, si f_λ es el minimizador en $\mathcal{H}_{\mathcal{X}}$ de $J_\lambda(f)$ definido en (2.14), entonces*

$$\hat{f}_\lambda = -\frac{\hat{\xi}}{\lambda} + \sum_{a=1}^{n}\sum_{i=1}^{p}\beta_{(a-1)p+i}\frac{\partial}{\partial x_i}k(\mathbf{x}_a, \cdot),$$

donde

$$\hat{\xi} := \frac{1}{n}\sum_{a=1}^{n}\sum_{i=1}^{p}\left(\frac{\partial}{\partial x_i}k(\mathbf{x}_a, \cdot)\frac{\partial}{\partial x_i}\log q_0(\mathbf{x}_a) + \frac{\partial^2}{\partial x_i^2}k(\mathbf{x}_a, \cdot)\right),$$

y $\boldsymbol{\beta} \in \mathbb{R}^{np}$ es la solución del sistema

$$(\mathbf{G} + n\lambda\mathbf{I}_{np})\boldsymbol{\beta} = \frac{\mathbf{h}}{\lambda},$$

con $\mathbf{G} \in \mathbb{R}^{np\times np}$ y $\mathbf{h} \in \mathbb{R}^{np}$ definidos como sigue:

$$(\mathbf{G})_{(a-1)p+i,(b-1)p+j} = \frac{\partial}{\partial x_i}\frac{\partial}{\partial x_{j+p}}k(\mathbf{x}_a, \mathbf{x}_b),$$

$$\mathbf{h}_{(a-1)p+i} = \frac{1}{n}\sum_{b=1}^{n}\sum_{j=1}^{p}\left(\frac{\partial}{\partial x_i}\frac{\partial}{\partial x_{j+p}}k(\mathbf{x}_a, \mathbf{x}_b)\frac{\partial}{\partial x_j}\log q_0(\mathbf{x}_b) + \frac{\partial}{\partial x_i}\frac{\partial^2}{\partial x_{j+p}^2}k(\mathbf{x}_a, \mathbf{x}_b)\right).$$

2.4 Comentarios de cierre de capítulo

En este capítulo presentamos los ingredientes fundamentales de modelado que usaremos luego para proponer nuevos métodos de reducción dimensional, con la finalidad de obtener una representación eficiente de la información predictiva. En particular, revisamos la familia exponencial clásica (EF) y presentamos una extensión basada en núcleos (KEF), la cual incrementa la flexibilidad de modelado manteniendo muchas de las propiedades interesantes de las EF. La ganancia de flexibilidad se debe fundamentalmente a la adopción de parámetros funcionales para caracterizar a la familia. Las propiedades de las KEF nos permitirán usar herramientas de modelado paramétrico para lograr representaciones eficientes de los datos, con niveles de expresividad semejantes a los de métodos no paramétricos. En el siguiente capítulo resumiremos resultados conocidos de reducción suficiente de dimensiones, completando con ello nuestra preparación para abordar los métodos propuestos en esta tesis.

CAPÍTULO 3

Reducción suficiente de dimensiones

Cuando intentamos establecer la relación entre un conjunto de variables predictoras $\mathbf{X} \in \mathbb{R}^p$ y una variable respuesta $Y \in \mathbb{R}$, la dimensión p es un factor relevante. A medida que p crece, es cada vez más difícil obtener representaciones gráficas que permitan visualizar, analizar e inferir acerca de la relación de Y como función de $\mathbf{X}$ y, en consecuencia, es cada vez más complicado proponer un modelo para $Y|\mathbf{X}$. Por esa razón, es de especial interés intentar definir una transformación $\mathbf{R}(\mathbf{X}) \in \mathbb{R}^q$, en lo posible con $q \ll p$, que además sea una buena representación para explotar de la mejor forma posible la información que tiene $\mathbf{X}$ sobre Y. En el mejor de los casos, $\mathbf{R}(\mathbf{X})$ será capaz de conservar toda la información predictiva que tiene $\mathbf{X}$ acerca de Y, lo cual permite afirmar que la reducción $\mathbf{R}(\mathbf{X})$ es suficiente.

En este capítulo repasaremos los conceptos básicos de esta temática, como punto de partida para estudiar en el próximo capítulo reducciones suficientes en la familia exponencial basada en núcleos (KEF) de la Definición 2.8. En la Sección 3.1 estudiaremos el concepto de reducción suficiente de dimensiones, su relación con el marco teórico de la suficiencia estadística y los principales resultados para poder obtener tales reducciones. Luego, en las Secciones 3.2 y 3.3 repasaremos los casos de las reducciones suficientes lineales y no lineales, respectivamente, comentando diferentes métodos representativos. Por último, otros métodos de reducción pertenecientes al estado del arte serán revisados en la Sección 3.4.

3.1 Definiciones básicas

De ahora en adelante, consideraremos la regresión de $Y \in \mathcal{Y} \subset \mathbb{R}$ en $\mathbf{X} \in \mathcal{X} \subset \mathbb{R}^p$. Denotaremos $\boldsymbol{\mu} = \mathbb{E}[\mathbf{X}]$ la esperanza de $\mathbf{X}$ y $\boldsymbol{\Sigma} = \operatorname{var}(\mathbf{X})$ la matriz de covarianza poblacional de $\mathbf{X}$. Por otra parte, para un conjunto $\mathcal{D}_{\mathbf{x}} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$, la media muestral $\bar{\mathbf{x}}$ y la matriz de covarianza muestral $\mathbf{S}$ están definidas por

$$\bar{\mathbf{x}} := \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i \qquad \text{y} \qquad \mathbf{S} := \frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^{\mathrm{T}},$$

mientras que $\hat{\boldsymbol{\Sigma}} := \frac{n-1}{n}\mathbf{S}$ es el MLE de $\boldsymbol{\Sigma}$.

Por último, usaremos $A =_{\mathrm{D}} B$ y $A \perp\!\!\!\perp B$ para indicar que A y B son v.a. idénticamente distribuidas e independientes, respectivamente.

Comenzaremos formalizando el concepto de *reducción suficiente de dimensiones* (SDR) a través de tres formas equivalentes. La idea básica es obtener una transformación $\mathbf{R}(\mathbf{X})$ de menor dimensión que $\mathbf{X}$, preservando toda la información predictiva que $\mathbf{X}$ tiene sobre Y. De este modo, en la regresión de Y en $\mathbf{X}$ será indistinto observar $\mathbf{X}$ o $\mathbf{R}(\mathbf{X})$.

Definición 3.1. [Cook, 2007] Una transformación $\mathbf{R} : \mathbb{R}^p \to \mathbb{R}^q$, con $q \leq p$, define una *reducción suficiente* (SDR) de $\mathbf{X}$ para la regresión de Y en $\mathbf{X}$ si satisface al menos una de las siguientes condiciones:

(I) (*Reducción inversa*) $\mathbf{X}|(Y, \mathbf{R}(\mathbf{X})) =_{\mathrm{D}} \mathbf{X}|\mathbf{R}(\mathbf{X})$.

(II) (*Reducción directa*) $Y|\mathbf{X} =_{\mathrm{D}} Y|\mathbf{R}(\mathbf{X})$.

(III) (*Reducción conjunta*) $\mathbf{X} \perp\!\!\!\perp Y|\mathbf{R}(\mathbf{X})$.

La condición (I) corresponde a la regresión inversa $\mathbf{X}|Y$, (II) a la regresión directa $Y|\mathbf{X}$, mientras que la condición (III) requiere la distribución conjunta de $(\mathbf{X}, Y)$. Más aún, las tres condiciones de la Definición 3.1 son equivalentes [Cook, 2007], por lo cual es posible obtener una reducción a partir de una de ellas y pasar a los otros enfoques de forma inmediata. Por ejemplo, podemos obtener una reducción suficiente a partir de

$\mathbf{X}|Y$ y pasar a la regresión directa $Y|\mathbf{X}$ o a la distribución conjunta, sin necesidad de especificar la distribución marginal de Y o la distribución condicional de $Y|\mathbf{X}$.

En particular, el enfoque de regresión inversa tiene la importante ventaja de simplificar el problema de la regresión multivariada de Y en $\mathbf{X}$ a trabajar con p regresiones univariadas, una por cada componente de $\mathbf{X}$ en Y, lo cual es más fácil de visualizar y modelar. La equivalencia entre las condiciones (I) y (II) de la Definición 3.1 permite establecer una conexión con el concepto de *estadístico suficiente* que enunciaremos a continuación, poniendo en evidencia que la idea de reducción suficiente está en armonía con la teoría existente que inició con el concepto clásico de suficiencia estadística en [Fisher, 1922].

Definición 3.2. [Casella and Berger, 2002, Definición 6.2.1] Sea $\boldsymbol{\theta} \in \mathbb{R}^m$ un parámetro de interés de la distribución de un conjunto de datos $\mathcal{D}_{\mathbf{x}} \subset \mathbb{R}^p$. Una transformación $\mathbf{T} : \mathbb{R}^p \to \mathbb{R}^q$ define un *estadístico suficiente* para $\boldsymbol{\theta}$ si se verifica

$$\mathcal{D}_{\mathbf{x}} \perp\!\!\!\perp \boldsymbol{\theta} | \mathbf{T}(\mathcal{D}_{\mathbf{x}}). \tag{3.1}$$

La Definición 3.2 puede interpretarse de la siguiente manera: si dos observaciones distintas $\mathbf{x}_1, \mathbf{x}_2 \in \mathcal{D}_{\mathbf{x}}$ verifican $\mathbf{T}(\mathbf{x}_1) = \mathbf{T}(\mathbf{x}_2)$, para el objetivo de inferir acerca de $\boldsymbol{\theta}$ es indistinto observar $\mathbf{x}_1$ ó $\mathbf{x}_2$. En consecuencia, $\mathbf{T}$ genera una partición de $\mathcal{X} \subset \mathbb{R}^p$ formada por los conjuntos $A_{\mathbf{t}} = \{\mathbf{x} \in \mathbb{R}^p : \mathbf{R}(\mathbf{x}) = \mathbf{t},\ \mathbf{t} \in \mathbf{T}(\mathcal{X})\}$; es decir, dicha partición está determinada por la relación de equivalencia $\mathcal{R}$ definida por

$$\mathbf{x}_1 \mathcal{R}\, \mathbf{x}_2 \Leftrightarrow \mathbf{T}(\mathbf{x}_1) = \mathbf{T}(\mathbf{x}_2). \tag{3.2}$$

Ahora bien, la conexión entre reducciones y estadísticos suficientes se establece como sigue: tomando $\mathcal{D}_{\mathbf{x}} = \mathbf{X}$, $\theta = Y$ y $\mathbf{T} = \mathbf{R}$ en la Definición 3.2, la expresión (3.1) resulta idéntica a la condición (III) de la Definición 3.1. Por lo tanto, podemos afirmar que $\mathbf{R}(\mathbf{X})$ es una SDR de $\mathbf{X}$ para la regresión de Y en $\mathbf{X}$ si y solo si $\mathbf{R}(\mathbf{X})$ es un estadístico suficiente al tratar a Y como un parámetro. Esta correspondencia entre ambos conceptos es fundamental y permite explotar los criterios preexistentes de búsqueda de estadísticos suficientes. Por supuesto, dichos criterios pueden reexpresarse en términos de reducciones

suficientes dentro el enfoque de regresión inversa $\mathbf{X}|Y$, en correspondencia con la condición (i) de la Definición 3.1. Esas versiones son las que expondremos en este capítulo, acorde a la temática de esta tesis. Por ejemplo, el siguiente teorema permite detectar reducciones suficientes a partir de la función de densidad condicional de $\mathbf{X}|Y$.

Teorema 3.3. (Teorema de Factorización) [Casella and Berger, 2002, Teorema 6.2.6] *Para todo $y \in \mathcal{Y}$, sea $p(\mathbf{x}|y)$ la función de densidad condicional de $\mathbf{X}|(Y = y)$. Una transformación $\mathbf{R} : \mathbb{R}^p \to \mathbb{R}^q$, con $q \leq p$, define una SDR de $\mathbf{X}$ para la regresión de Y en $\mathbf{X}$ si y solo si existen funciones $g(\mathbf{t}, y)$ y $h(\mathbf{x})$ tales que, para todo $\mathbf{x} \in \mathcal{X}$ y todo $y \in \mathcal{Y}$, se verifica*

$$p(\mathbf{x}|y) = g(\mathbf{R}(\mathbf{x}), y)\, h(\mathbf{x}).$$

Una consecuencia directa del Teorema 3.3 es que la suficiencia se mantiene bajo transformaciones inyectivas, lo cual permite establecer nuevas reducciones suficientes a partir de una dada y, por consiguiente, nos da una noción para definir una equivalencia entre reducciones suficientes. Formalizaremos esto con el siguiente resultado, cuya demostración se encuentra en el Anexo B.1.

Corolario 3.4. *Sea $\mathbf{R} : \mathbb{R}^p \to \mathbb{R}^q$ una SDR de $\mathbf{X}$ para la regresión de Y en $\mathbf{X}$. Si $\mathbf{S} : \mathbb{R}^q \to \mathbb{R}^m$ es una función inyectiva, entonces $\mathbf{S} \circ \mathbf{R}$ también es una SDR de $\mathbf{X}$ para la regresión de Y en $\mathbf{X}$.*

Bajo las condiciones del Corolario 3.4, tanto $\mathbf{R}(\mathbf{X})$ como $(\mathbf{S} \circ \mathbf{R})(\mathbf{X})$ determinan en $\mathcal{X} \subset \mathbb{R}^p$ la misma partición. En efecto, por inyectividad de $\mathbf{S}$ y de (3.2), podemos escribir

$$\mathbf{x}_1 \mathcal{R}\, \mathbf{x}_2 \Leftrightarrow \mathbf{R}(\mathbf{x}_1) = \mathbf{R}(\mathbf{x}_2) \Leftrightarrow (\mathbf{S} \circ \mathbf{R})(\mathbf{x}_1) = (\mathbf{S} \circ \mathbf{R})(\mathbf{x}_2).$$

Además, si ambas reducen $\mathbf{X}$ a la misma dimensión ($q = m$) podríamos decir que tienen el mismo valor en términos de suficiencia. Esto motiva la siguiente definición.

Definición 3.5. Sean $\mathbf{R}_1 : \mathbb{R}^p \to \mathbb{R}^q$ y $\mathbf{R}_2 : \mathbb{R}^p \to \mathbb{R}^q$, ambas SDR de $\mathbf{X}$ para la regresión de Y en $\mathbf{X}$. Diremos que $\mathbf{R}_1(\mathbf{X})$ y $\mathbf{R}_2(\mathbf{X})$ son *reducciones suficientes equivalentes* si existe una función inyectiva $\mathbf{S} : \mathbb{R}^q \to \mathbb{R}^q$ tal que $\mathbf{R}_1(\mathbf{X}) = (\mathbf{S} \circ \mathbf{R}_2)(\mathbf{X})$; o lo que es lo mismo, si para todo $\mathbf{x}_1, \mathbf{x}_2 \in \mathcal{X}$ se verifica

$$\mathbf{R}_1(\mathbf{x}_1) = \mathbf{R}_1(\mathbf{x}_2) \Leftrightarrow \mathbf{R}_2(\mathbf{x}_1) = \mathbf{R}_2(\mathbf{x}_2).$$

La situación óptima sería obtener una SDR que genere la partición más gruesa de $\mathcal{X} \subset \mathbb{R}^p$ y además minimice el valor de la dimensión q. El valor óptimo de q, conocido como *dimensión estructural* de $\mathbf{X}$ para la regresión de Y en $\mathbf{X}$, se suele denotar con d. Por ahora nada asegura la existencia de una SDR con dichas características, pero en caso de existir, está caracterizada por la siguiente definición.

Definición 3.6. Una transformación $\mathbf{R}_0 : \mathbb{R}^p \to \mathbb{R}^d$, con $d \leq p$, se denomina *reducción suficiente minimal* de $\mathbf{X}$ para la regresión de Y en $\mathbf{X}$ si: (i) es una SDR de $\mathbf{X}$, y (ii) para cualquier otra SDR de $\mathbf{X}$, a saber $\mathbf{R} : \mathbb{R}^p \to \mathbb{R}^q$, existe una función $\mathbf{S} : \mathbb{R}^q \to \mathbb{R}^d$ tal que

$$\mathbf{R}_0(\mathbf{X}) = (\mathbf{S} \circ \mathbf{R})(\mathbf{X}).$$

Como es de esperar, el concepto de SDR minimal tiene su par equivalente para el caso de estadísticos suficientes. En este contexto, el Teorema de Lehmann-Scheffé es de gran utilidad para encontrar estadísticos suficientes minimales. A continuación, lo enunciamos en términos de reducciones suficientes.

Teorema 3.7. (Teorema de Lehmann-Scheffé) [Casella and Berger, 2002, Teorema 6.2.13] *Para todo $y \in \mathcal{Y}$, sea $p(\mathbf{x}|y)$ la función de densidad condicional de $\mathbf{X}|(Y = y)$. Una transformación $\mathbf{R} : \mathbb{R}^p \to \mathbb{R}^d$, con $d \leq p$, es una SDR minimal de $\mathbf{X}$ para la regresión de Y en $\mathbf{X}$ si, para todo $\mathbf{x}_1, \mathbf{x}_2 \in \mathcal{X}$, se verifica*

$$\frac{p(\mathbf{x}_1|y)}{p(\mathbf{x}_2|y)} \text{ es constante como función de } y \Leftrightarrow \mathbf{R}(\mathbf{x}_1) = \mathbf{R}(\mathbf{x}_2).$$

Ya estamos en condiciones de explorar los distintos enfoques y métodos para obtener reducciones suficientes. Antes de eso, haremos un par de observaciones:

- Es trivial que $\mathbf{R}(\mathbf{X}) = \mathbf{X}$ verifica las condiciones de la Definición 3.1 y, en consecuencia, es una SDR de $\mathbf{X}$ para la regresión de Y en $\mathbf{X}$. Más aún, cualquier función inyectiva de $\mathbf{X}$ también es una SDR, aunque en ningún caso estaríamos realmente reduciendo la dimensión de $\mathbf{X}$. De hecho, puede ocurrir que no sea posible reducir, lo cual implicaría que todas las SDR de $\mathbf{X}$ sean equivalentes a $\mathbf{R}(\mathbf{X}) = \mathbf{X}$.
- Una SDR de $\mathbf{X}$ para la regresión de Y en $\mathbf{X}$ puede ser o no una función lineal de $\mathbf{X}$. La linealidad es muchas veces impuesta para facilitar el análisis, o bien es una consecuencia directa de adoptar ciertos modelos para la regresión inversa $\mathbf{X}|Y$. Por esa razón, dividiremos el repaso de algunas de las metodologías existentes en casos de SDR lineal (Subsección 3.2) y SDR no lineal (Subsección 3.3).

3.2 Reducción suficiente lineal

La mayoría de los métodos de SDR buscan transformaciones lineales de la variable predictora $\mathbf{X} \in \mathbb{R}^p$. Es decir, tienen como objetivo encontrar una matriz $\mathbf{\Gamma} \in \mathbb{R}^{p\times q}$, con $q \leq p$, tal que $\mathbf{R}(\mathbf{X}) = \mathbf{\Gamma}^{\mathrm{T}}\mathbf{X}$ sea una SDR de $\mathbf{X}$ para la regresión de Y en $\mathbf{X}$. Para este caso, la condición de regresión directa (II) de la Definición 3.1 se escribe

$$Y|\mathbf{X} =_{\mathrm{D}} Y|\mathbf{\Gamma}^{\mathrm{T}}\mathbf{X}. \tag{3.3}$$

Del Corolario 3.4 y de la Definición 3.5 se deduce que, si $\mathbf{A} \in \mathbb{R}^{q\times q}$ es una matriz invertible y (3.3) es cierto, entonces $\tilde{\mathbf{R}}(\mathbf{X}) = \mathbf{A}\mathbf{\Gamma}^{\mathrm{T}}\mathbf{X}$ también es una SDR de $\mathbf{X}$ para la regresión de Y en $\mathbf{X}$. Además, $\tilde{\mathbf{R}}(\mathbf{X})$ es equivalente a $\mathbf{R}(\mathbf{X})$ y, a la vez, $\operatorname{span}\mathbf{\Gamma} = \operatorname{span}\mathbf{\Gamma}\mathbf{A}^{\mathrm{T}}$, lo cual significa que el espacio generado por las columnas de $\mathbf{\Gamma}$ es invariante por transformaciones lineales biyectivas. Por lo tanto, el interés no radica en $\mathbf{\Gamma}$ en sí mismo, sino en el subespacio que genera, por lo cual es conveniente reescribir (3.3) como

$$Y|\mathbf{X} =_{\mathrm{D}} Y|\mathbf{P}_{\mathbf{\Gamma}}\mathbf{X},$$

donde $\mathbf{P}_{\boldsymbol{\Gamma}}$ es la matriz de proyección ortogonal sobre $\operatorname{span}\boldsymbol{\Gamma}$. Utilizaremos la condición equivalente (II) de la Definición 3.1 para el siguiente concepto.

Definición 3.8. [Cook, 1998, Sección 6.2] Un subespacio $\mathcal{S} \subset \mathbb{R}^p$ es un *subespacio de reducción suficiente* (DRS) para la regresión de Y en $\mathbf{X}$ si se verifica

$$\mathbf{X} \perp\!\!\!\perp Y | \mathbf{P}_{\mathcal{S}}\mathbf{X}, \tag{3.4}$$

donde $\mathbf{P}_{\mathcal{S}}$ es la matriz de proyección ortogonal sobre $\mathcal{S}$.

Claramente estamos interesados en obtener un subespacio $\mathcal{S}$ que verifique (3.4) y cuya dimensión sea lo más chica posible. En otras palabras, nos gustaría hallar un DRS que esté contenido en todos los demás DRS. La existencia de dicho subespacio no está asegurada pero, en caso de existir, queda caracterizado de la siguiente manera.

Definición 3.9. [Cook, 1998, Sección 6.3] Un DRS $\mathcal{S}_{Y|\mathbf{X}}$ para la regresión de Y en $\mathbf{X}$ se denomina *subespacio central* (CS) si, para cualquier otro DRS $\mathcal{S}$, se verifica $\mathcal{S}_{Y|\mathbf{X}} \subset \mathcal{S}$.

Es evidente que el CS para la regresión de Y en $\mathbf{X}$ existe si y solo si la intersección $\cap\mathcal{S}$ de todos los DRS es también un DRS, en cuyo caso $\mathcal{S}_{Y|\mathbf{X}} = \cap\mathcal{S}$. Ahora bien, la intersección $\cap\mathcal{S}$ no necesariamente es un DRS y, en consecuencia, $\mathcal{S}_{Y|\mathbf{X}}$ puede no existir. No obstante, al escribir $\mathcal{S}_{Y|\mathbf{X}}$ asumiremos que sí existe, en cuyo caso $\mathcal{S}_{Y|\mathbf{X}}$ es único [Cook, 1998, Proposición 6.2] y la dimensión estructural es $d = \dim(\mathcal{S}_{Y|\mathbf{X}})$. Continuamos con la siguiente propiedad del CS, que suele ser de mucha utilidad.

Proposición 3.10. [Cook, 1998, Proposición 6.3] *Sea* $\tilde{\mathbf{X}} = \mathbf{A}^{\mathrm{T}}\mathbf{X}$, *con* $\mathbf{A} \in \mathbb{R}^{p\times p}$ *una matriz de rango completo. Se verifica*

$$\mathcal{S}_{Y|\tilde{\mathbf{X}}} = \mathbf{A}^{-1}\mathcal{S}_{Y|\mathbf{X}}.$$

En particular, para la estandarización $\mathbf{Z} := \boldsymbol{\Sigma}^{-1/2}(\mathbf{X} - \boldsymbol{\mu})$ *resulta*

$$\mathcal{S}_{Y|\mathbf{Z}} = \boldsymbol{\Sigma}^{1/2}\mathcal{S}_{Y|\mathbf{X}}. \tag{3.5}$$

La relación (3.5) permite utilizar $\mathbf{Z}$ cuando se considere conveniente, y luego obtener $\mathcal{S}_{Y|\mathbf{X}}$ a partir de $\mathcal{S}_{Y|\mathbf{Z}}$ mediante

$$\mathcal{S}_{Y|\mathbf{X}} = \boldsymbol{\Sigma}^{-1/2}\mathcal{S}_{Y|\mathbf{Z}}. \tag{3.6}$$

Ahora bien, los métodos de reducción lineal consideran el problema de hallar $\boldsymbol{\alpha} \in \mathbb{R}^{p\times d}$ tal que

$$\mathbf{X} \perp\!\!\!\perp Y|\mathbf{P}_{\boldsymbol{\alpha}}\mathbf{X} \qquad \text{y} \qquad \operatorname{span}\boldsymbol{\alpha} = \mathcal{S}_{Y|\mathbf{X}}. \tag{3.7}$$

Bajo (3.7), una SDR minimal de $\mathbf{X}$ para la regresión de Y en $\mathbf{X}$ es

$$\mathbf{R}(\mathbf{X}) = \boldsymbol{\alpha}^{\mathrm{T}}\mathbf{X}.$$

Por lo tanto, el objetivo principal es estimar una base cualquiera $\boldsymbol{\alpha}$ para el subespacio central $\mathcal{S}_{Y|\mathbf{X}}$, basándose en ciertas suposiciones y/o modelos que permitan caracterizar $\mathcal{S}_{Y|\mathbf{X}}$, y luego proponer una metodología para estimarlo. El modelo más simple para la regresión directa $Y|\mathbf{X}$ es la *regresión lineal múltiple* (MLR), cuya formulación es

$$y = \boldsymbol{\alpha}^{\mathrm{T}}\mathbf{X} + \varepsilon, \tag{3.8}$$

donde ε es un error aleatorio y $\boldsymbol{\alpha}$ es estimado mediante la técnica clásica de *mínimos cuadrados* (OLS). Pese a las fuertes suposiciones que realiza, MLR sigue estando vigente debido a su practicidad y fácil interpretabilidad. Uno de los principales inconvenientes que atenta contra OLS es la posible existencia de alta correlación entre ciertas variables predictoras, lo cual es propio de problemas con un alto número de variables. Para estos casos, una buena alternativa es utilizar *mínimos cuadrados parciales* (PLS), el cual en particular suele tener buen rendimiento en escenarios de alta dimensionalidad. El origen de PLS se remonta al trabajo de Wold [1966], quien luego propuso el algoritmo NIPALS [Wold, 1975]. Más adelante, de Jong [1993] propuso un eficiente algoritmo para PLS, denominado SIMPLS, el cual es probablemente el más utilizado en la actualidad.

El modelo (3.8) puede ser reemplazado por una formulación más general:

$$y = f(\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{X}) + \varepsilon. \tag{3.9}$$

Li [1991] analizó (3.9) sin realizar suposiciones sobre la función f y el error ε. Además, sugirió tratar el problema mediante el enfoque de regresión inversa $\mathbf{X}|Y$, bajo el cual formuló su método *Sliced Inverse Regression* (SIR). En dicho trabajo, centra la atención en la curva de regresión $\mathbb{E}[\mathbf{Z}|Y]$ y prueba que, bajo ciertas condiciones, ésta queda contenida en el subespacio central $\mathcal{S}_{Y|\mathbf{Z}}$. SIR es un método basado en el primer momento de $\mathbf{X}|Y$ y es el precursor tanto del enfoque de regresión inversa como de la metodología de SDR basada en momentos. Repasaremos SIR con más detalle en la Subsección 3.2.1.

Inmediatamente después de SIR surge un método basado en el segundo momento de $\mathbf{X}|Y$, denominado *Sliced Average Variance Estimation* (SAVE) [Cook and Weisberg, 1991]. Allí proponen estimar $\mathcal{S}_{Y|\mathbf{Z}}$ a partir de una función de la varianza condicional $\text{var}(\mathbf{Z}|Y)$. Tanto SIR como SAVE impulsaron el análisis de funciones de los momentos de $\mathbf{X}|Y$, surgiendo así diferentes métodos que, bajo ciertas condiciones, detectan direcciones de reducción contenidas en $\mathcal{S}_{Y|\mathbf{X}}$. Algunos ejemplos son *Directional Regression* (DR) [Li and Wang, 2007] y *Parametric Inverse Regression* (PIR) [Bura and Cook, 2001]. Por otra parte, Li también propuso un método basado en el segundo momento pero de la regresión directa $Y|\mathbf{X}$, denominado *Principal Hessian Directions* (pHd) [Li, 1992], que explota el hecho de que la matriz Hessiana $\mathbf{H}(\mathbf{X})$ de la función de regresión $\mathbb{E}[Y|\mathbf{X}]$, definida por $\mathbf{H}_{ij} := (\partial^2/\partial x_i \partial x_j \mathbb{E}[Y|\mathbf{X}])$, se degenera a lo largo de cualquier dirección ortogonal al subespacio de reducción para $\mathbb{E}[Y|\mathbf{X}]$.

Los métodos basados en momentos tienen ciertas limitaciones. Por ejemplo, además de realizar suposiciones sobre los momentos de $\mathbf{X}|Y$, SIR tiende a estimar con frecuencia un subespacio propio de $\mathcal{S}_{Y|\mathbf{X}}$, mientras que SAVE, aunque aborda esta limitación, posee una capacidad menor que SIR para detectar tendencias lineales. Otro caso es pHd, que asume que $\mathbf{X}$ tiene distribución normal. No obstante, estos métodos son fáciles de aplicar y suelen ser muy utilizados en diferentes contextos, dando lugar a numerosas variantes.

Otra alternativa dentro del enfoque de regresión inversa es proponer algún modelo paramétrico para $\mathbf{X}|Y$, lo cual da lugar a la metodología de SDR basada en modelos. En particular, si las distribuciones condicionales de $\mathbf{X}|Y$ se modelan en una EF, en [Bura et al., 2016] hallaron una fórmula cerrada para una SDR minimal de $\mathbf{X}$ para la regresión

de Y en $\mathbf{X}$, que si bien es cierto no es lineal en $\mathbf{X}$, sí lo es en su estadístico suficiente $\mathbf{T}(\mathbf{X})$. Dado que en general esto se traduce en una transformación no lineal de $\mathbf{X}$, repasaremos dicha metodología dentro de la Sección 3.3, más precisamente en la Subsección 3.3.1.

Dentro del enfoque de reducción lineal basada en modelos, una opción muy estudiada es analizar $\mathbf{X}|Y$ bajo normalidad y estimar reducciones suficientes mediante estimadores de verosimilitud. En general, se plantea el modelo $\mathbf{X}|(Y = y) \sim \mathcal{N}_p(\boldsymbol{\mu}_y, \boldsymbol{\Delta}_y)$, con $\boldsymbol{\mu}_y \in \mathbb{R}^p$ y $\boldsymbol{\Delta}_y \in \mathbb{R}^{p\times p}$ s.d.p. [ver Cook et al., 2011]. Dicho modelo puede ser expresado de manera equivalente como

$$\mathbf{X}_y := \mathbf{X}|(Y = y) = \boldsymbol{\mu}_y + \boldsymbol{\epsilon}, \tag{3.10}$$

donde $\boldsymbol{\epsilon} \sim \mathcal{N}_p(\mathbf{0}, \boldsymbol{\Delta}_y)$ se denomina *error* y puede o no depender de y (si no depende de y, se escribe simplemente $\boldsymbol{\Delta}$). Con el objetivo de construir un MLE de $\mathcal{S}_{Y|\mathbf{X}}$, se analizan diferentes estructuras para $\boldsymbol{\mu}_y$ en (3.10) y se consideran diferentes casos de $\boldsymbol{\Delta}_y$, facilitando la construcción del estimador. Algunos ejemplos son *Principal Fitted Components* (PFC) [Cook and Forzani, 2008], el cual repasaremos en la Subsección 3.2.2, y *Likelihood Acquired Directions* (LAD) [Cook and Forzani, 2009]. Cabe destacar que en [Cook et al., 2011] se brinda una implementación unificada de estos métodos en `MATLAB`, mientras que una adaptación a `R` fue publicada luego en [Adragni and Raim, 2014].

De los métodos de reducción lineal arriba mencionados, en la parte experimental que desarrollaremos en el Capítulo 5 utilizaremos el método SIR, basado en momentos, y el método PFC, basado en modelos. La elección se basa en lo importante que son dichos métodos dentro del estado del arte, su fácil implementación y su buen rendimiento en escenarios de baja dimensionalidad. A continuación, haremos una breve descripción de ellos para entender más en detalle en qué consisten.

3.2.1 Sliced Inverse Regression (SIR) [Li, 1991]

Li [1991] probó en primer lugar que, bajo condiciones no estrictas, la curva de regresión inversa $\mathbb{E}[\mathbf{X}|Y]$ en $\mathbb{R}^p$ está contenida en un subespacio de dimensión d, directamente relacionado con las columnas de $\boldsymbol{\alpha}$, tal como se establece a continuación.

Teorema 3.11. [Li, 1991, Teorema 3.1] *Bajo (3.7), si la esperanza condicional* $\mathbb{E}[\mathbf{X}|\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{X}]$ *es lineal en* $\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{X}$*, entonces* $\mathbb{E}[\mathbf{X}|Y] - \boldsymbol{\mu}$ *está contenida en el subespacio* $\operatorname{span}(\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{\Sigma})$.

En particular, para la estandarización de $\mathbf{X}$ se obtiene el siguiente resultado, el cual permite identificar una parte del subespacio central $\mathcal{S}_{Y|\mathbf{Z}}$.

Corolario 3.12. [Li, 1991, Corolario 3.1] *Bajo las condiciones del Teorema 3.11, la curva de regresión estandarizada* $\mathbb{E}[\mathbf{Z}|Y]$ *está contenida en el subespacio* $\mathcal{S}_{Y|\mathbf{Z}}$.

La *condición de linealidad* del resultado anterior es equivalente a $\mathbb{E}[\mathbf{Z}|\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{Z}] = \mathbf{P}_{\mathcal{S}_{Y|\mathbf{Z}}}\mathbf{Z}$. Esta impone una restricción en la distribución marginal de $\mathbf{X}$, la cual se cumple si $\mathbf{X}$ tiene distribución elípticamente simétrica; por ejemplo, la distribución Normal Multivariada.

En virtud del Corolario 3.12, la matriz de covarianza $\operatorname{var}(\mathbb{E}[\mathbf{Z}|Y])$ se degenera en cualquier dirección ortogonal a $\mathcal{S}_{Y|\mathbf{Z}}$. En consecuencia, sus autovectores están directamente relacionados al CS.

Corolario 3.13. [Li, 1991] *Bajo las condiciones del Teorema 3.11, los autovectores correspondientes a los d autovalores más grandes de la matriz*

$$\mathbf{K}_{\mathrm{SIR}} := \operatorname{var}(\mathbb{E}[\mathbf{Z}|Y]) \tag{3.11}$$

pertenecen al subespacio $\mathcal{S}_{Y|\mathbf{Z}}$.

En virtud del Corolario 3.13 y de la relación (3.6), surge el método de estimación de $\mathcal{S}_{Y|\mathbf{X}}$ denominado *Sliced Inverse Regression* (SIR), el cual detallamos a continuación.

Método de estimación **SIR**

---

1. Estandarizar los datos mediante

$$\mathbf{z}_i = \hat{\boldsymbol{\Sigma}}^{-1/2}(\mathbf{x}_i - \bar{\mathbf{x}}).$$

2. Particionar el rango de Y en H *slices*; esto es, seleccionar $\{I_1, \ldots, I_H\}$ disjuntos dos a dos tales que $\mathcal{Y} = I_1 \cup \cdots \cup I_H$. La proporción de datos en el *slice* I_h es

$$\hat{p}_h = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}\left(y_i \in I_h\right).$$

3. Dentro de cada *slice*, calcular la media muestral de los datos estandarizados mediante

$$\hat{\mathbf{m}}_h = \frac{1}{n\hat{p}_h} \sum_{\tilde{y}_i = h} \mathbf{z}_i.$$

4. Calcular los autovectores y autovalores del estimador de $\mathbf{K}_{\text{SIR}}$ de (3.11) dado por

$$\hat{\mathbf{K}}_{\text{SIR}} = \sum_{h=1}^{H} \hat{p}_h \hat{\mathbf{m}}_h \hat{\mathbf{m}}_h^{\mathrm{T}}.$$

5. Sea $\hat{\mathbf{B}} \in \mathbb{R}^{p \times d}$ la matriz cuyas columnas son los autovectores correspondientes a los d autovalores más grandes de $\hat{\mathbf{K}}_{\text{SIR}}$. El estimador SIR de $\boldsymbol{\alpha}$ es

$$\hat{\boldsymbol{\alpha}} = \hat{\boldsymbol{\Sigma}}^{-1/2} \hat{\mathbf{B}}.$$

Es importante remarcar que SIR no proporciona una estimación exhaustiva de $\mathcal{S}_{Y|\mathbf{X}}$, sino que en general estima un subespacio propio de $\mathcal{S}_{Y|\mathbf{X}}$. En efecto, $\tilde{\mathbf{K}}_{\text{SIR}} := \text{var}(\mathbb{E}[\mathbf{Z}|\tilde{Y}])$, donde $\tilde{Y}$ es una discretización de Y, verifica $\text{span}\left(\tilde{\mathbf{K}}_{\text{SIR}}\right) \subset \mathcal{S}_{Y|\mathbf{Z}}$. Esto implica que SIR suele detectar solo algunas direcciones de reducción, especialmente aquellas relacionadas con tendencias lineales de $\mathbb{E}\left[Y|\mathbf{X}\right]$.

3.2.2 Principal Fitted Components (PFC) [Cook and Forzani, 2008]

Bajo el modelo (3.10), Cook [2007] desarrolló el caso especial de errores isotrópicos; esto es, $\boldsymbol{\epsilon} \sim \mathcal{N}_p(\mathbf{0}, \boldsymbol{\Delta})$ con $\boldsymbol{\Delta} = \sigma^2 \mathbf{I}_p$. Si $\boldsymbol{\Gamma} \in \mathbb{R}^{p \times d}$ es una matriz semi-ortogonal[1] cuyas columnas forman una base para $\text{span}\left\{\boldsymbol{\mu}_y - \boldsymbol{\mu} : y \in \mathcal{Y}\right\}$, podemos reescribir (3.10) como

$$\mathbf{X}_y = \boldsymbol{\mu} + \boldsymbol{\Gamma}\boldsymbol{\nu}_y + \boldsymbol{\epsilon}, \tag{3.12}$$

[1]Una matriz $\mathbf{A} \in \mathbb{R}^{n \times m}$ es semi-ortogonal si $\mathbf{A}^{\mathrm{T}}\mathbf{A} = \mathbf{I}_d$.

con $\boldsymbol{\epsilon} \sim \mathcal{N}_p(\mathbf{0}, \sigma^2\mathbf{I}_p)$ y $\boldsymbol{\nu}_y \in \mathbb{R}^d$ un vector desconocido. Resulta que $\mathbf{R}(\mathbf{X}) = \boldsymbol{\Gamma}^{\mathrm{T}}\mathbf{X}$ es una SDR de $\mathbf{X}$ para la regresión de Y en $\mathbf{X}$ [Cook, 2007, Proposición 1] y, por lo tanto, el objetivo es estimar el DRS $\mathcal{S}_{\boldsymbol{\Gamma}} := \operatorname{span}\boldsymbol{\Gamma}$.

El modelo (3.12) se conoce como *modelo PC*, debido a que el MLE de $\mathcal{S}_{\boldsymbol{\Gamma}}$ coincide con el tradicional *análisis de componentes principales* (PCA). Es decir, se estima $\mathcal{S}_{\boldsymbol{\Gamma}}$ mediante el subespacio generado por los autovectores correspondientes a los d autovalores más grandes de $\hat{\boldsymbol{\Sigma}}$. El principal inconveniente de este modelo es que solamente trabaja con la variable predictora $\mathbf{X}$ y no hace uso de la variable respuesta Y.

Para incorporar la información de Y, Cook [2007] propone una variante de (3.12), la cual consiste en modelar $\boldsymbol{\nu}_y \in \mathbb{R}^d$ como

$$\boldsymbol{\nu}_y = \boldsymbol{\beta}(\mathbf{f}_y - \mathbb{E}[\mathbf{f}_Y]), \tag{3.13}$$

donde $\mathbf{f}_y \in \mathbb{R}^r$ es una función vectorial de y que se asume conocida, con elementos linealmente independientes, y la matriz $\boldsymbol{\beta} \in \mathbb{R}^{d\times r}$, con $d \leq \min\{r, p\}$, no tiene restricciones de rango. Si bien aún se considera la estructura isotrópica del error dada por $\boldsymbol{\epsilon} \sim \mathcal{N}_p(\mathbf{0}, \sigma^2\mathbf{I}_p)$, luego en [Cook and Forzani, 2008] se extiende el análisis a errores con varianza constante; esto es, $\boldsymbol{\epsilon} \sim \mathcal{N}_p(\mathbf{0}, \boldsymbol{\Delta})$ con $\boldsymbol{\Delta} > 0$. De esta manera, a partir de reemplazar (3.13) en (3.12), se obtiene el *modelo PFC*:

$$\mathbf{X}_y = \tilde{\boldsymbol{\mu}} + \boldsymbol{\Gamma}\boldsymbol{\beta}\mathbf{f}_y + \boldsymbol{\epsilon}, \tag{3.14}$$

donde $\tilde{\boldsymbol{\mu}} = \boldsymbol{\mu} - \boldsymbol{\Gamma}\boldsymbol{\beta}\mathbb{E}[\mathbf{f}_y]$. Observar que, bajo (3.14), cada coordenada de $\mathbf{X}_y$ sigue un modelo de regresión lineal con predictor $\mathbf{f}_y$, lo cual permite utilizar gráficos de dichas coordenadas para obtener información acerca de cómo elegir convenientemente cada función $\mathbf{f}_y$. En esta situación descrita, en [Cook and Forzani, 2008] obtienen un resultado de reducción suficiente.

Teorema 3.14. [Cook and Forzani, 2008, Teorema 2.1] *Bajo el modelo (3.14), la transformación* $\mathbf{R} : \mathbb{R}^p \to \mathbb{R}^d$ *definida por*

$$\mathbf{R}(\mathbf{X}) = \boldsymbol{\Gamma}^{\mathrm{T}}\boldsymbol{\Delta}^{-1}\mathbf{X}$$

es una SDR minimal de $\mathbf{X}$ *para la regresión de* Y *en* $\mathbf{X}$.

Como consecuencia del Teorema 3.14, para el modelo PFC se verifica

$$\mathcal{S}_{Y|\mathbf{X}} = \operatorname{span} \boldsymbol{\Delta}^{-1}\boldsymbol{\Gamma}. \tag{3.15}$$

Observar que si $\boldsymbol{\Delta} = \sigma^2 \mathbf{I}_p$, la expresión (3.15) se simplifica a $\mathcal{S}_{Y|\mathbf{X}} = \operatorname{span} \boldsymbol{\Gamma}$. Este caso particular se conoce como *modelo PFC isotrópico*.

De (3.15), queda claro que el objetivo es estimar $\mathcal{S}_{Y|\mathbf{X}}$ mediante el MLE de $\boldsymbol{\Delta}^{-1}\boldsymbol{\Gamma}$. En [Cook and Forzani, 2008] se analizan también otras estructuras para $\boldsymbol{\Delta}$; por ejemplo, $\boldsymbol{\Delta}$ una matriz diagonal. A continuación, presentamos el MLE para el caso general $\boldsymbol{\Delta} > 0$.

Método de estimación **PFC**

Sea $\hat{\boldsymbol{\Sigma}}_{\text{fit}}$ la matriz de covarianza muestral de los vectores ajustados de la regresión lineal multivariada de $\mathbf{X}_y$ en $\mathbf{f}_y$, incluido un término independiente, y sea $\hat{\boldsymbol{\Sigma}}_{\text{res}}$ la matriz de covarianza de los vectores residuales de $\mathbf{X}_y$ en $\mathbf{f}_y$. Se verifica $\hat{\boldsymbol{\Sigma}}_{\text{res}} = \hat{\boldsymbol{\Sigma}} - \hat{\boldsymbol{\Sigma}}_{\text{fit}}$.

El MLE de $\mathcal{S}_{Y|\mathbf{X}}$ es el subespacio generado por $\hat{\boldsymbol{\Sigma}}_{\text{res}}^{-1/2}\hat{\mathbf{V}}$, donde $\hat{\mathbf{V}} \in \mathbb{R}^{p\times d}$ es la matriz cuyas columnas son los autovectores correspondientes a los d autovalores más grandes de $\hat{\boldsymbol{\Sigma}}_{\text{res}}^{-1/2}\hat{\boldsymbol{\Sigma}}\hat{\boldsymbol{\Sigma}}_{\text{res}}^{-1/2}$. En otras palabras, $\hat{\mathbf{V}}$ es la matriz resultante de aplicar PCA sobre $\hat{\boldsymbol{\Sigma}}_{\text{res}}^{-1/2}\mathbf{X}$.

En [Cook and Forzani, 2008, Corolario 3.4] se presentan también otras formas alternativas del MLE de $\mathcal{S}_{Y|\mathbf{X}}$. Una de ellas permite probar que, cuando Y es categórica y $\mathbf{f}_y$ se construye como un vector indicatriz para las clases, el estimador PFC de $\mathcal{S}_{Y|\mathbf{X}}$ es equivalente al estimador del método SIR.

3.3 Reducción suficiente no lineal

En esta sección repasaremos el estado del arte en lo que se refiere a SDR no lineales de $\mathbf{X}$, lo cual corresponde al problema general

$$Y \perp\!\!\!\perp \mathbf{X}|\mathbf{R}(\mathbf{X}), \tag{3.16}$$

con $\mathbf{R} : \mathbb{R}^p \to \mathbb{R}^q$ una transformación no lineal y $q \leq p$.

En cuanto a modelos paramétricos, cuando $\mathbf{X}|Y$ se modela en la familia exponencial $\mathcal{P}_{\text{fin}}$ dada por (2.5), en [Bura et al., 2016] identificaron una expresión cerrada para una SDR de $\mathbf{X}$ para la regresión de Y en $\mathbf{X}$, que a pesar de no ser lineal en $\mathbf{X}$ sí lo es en el estadístico suficiente $\mathbf{T}(\mathbf{X})$ de la familia exponencial. Su método de estimación, denominado EF-DR, lo veremos más en detalle en la Subsección 3.3.1.

Otra estrategia para estudiar el problema (3.16) es transformar la variable predictora $\mathbf{X}$ a través de una función $\phi : \mathbb{R}^p \to \mathcal{H}_\mathcal{X}$, donde $\mathcal{H}_\mathcal{X}$ es un espacio de Hilbert, y trabajar luego con $\phi(\mathbf{X})$. Es decir, se estudia en $\mathcal{H}_\mathcal{X}$ el problema de hallar $\tilde{\mathbf{\Gamma}}$ tal que

$$Y \perp\!\!\!\perp \mathbf{X} | \tilde{\mathbf{\Gamma}} \boldsymbol{\phi}(\mathbf{X}). \tag{3.17}$$

Observar que, dependiendo de la naturaleza de $\mathcal{H}_\mathcal{X}$, $\tilde{\mathbf{\Gamma}}$ puede ser una matriz o un operador funcional. La expresión (3.17) hace pensar en la aplicación de métodos de reducción lineal sobre la variable transformada $\phi(\mathbf{X})$, que luego posiblemente se traducen en reducciones no lineales de la variable original $\mathbf{X}$.

A la hora de elegir la transformación $\phi(\mathbf{X})$, los núcleos reproductores son una opción muy apropiada. Si $k : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}$ es un núcleo reproductor de un espacio de Hilbert $\mathcal{H}_\mathcal{X}$, se considera el mapeo característico $\phi : \mathbb{R}^p \to \mathcal{H}_\mathcal{X}$; es decir, $\phi(\mathbf{X}) := k(\mathbf{X}, \cdot)$. El éxito de esta herramienta se debe a que, aunque el RKHS $\mathcal{H}_\mathcal{X}$ generalmente es de dimensión infinita, los métodos basados en núcleos no requieren conocerlo explícitamente. De hecho, la matriz de Gram $\mathbf{K}_\mathbf{x} \in \mathbb{R}^{n \times n}$ asociada a un conjunto de datos $\mathcal{D}_\mathbf{x}$, definida por

$$(\mathbf{K}_\mathbf{x})_{ij} := k(\mathbf{x}_i, \mathbf{x}_j),$$

es todo lo que necesitamos a la hora de estimar las reducciones. Esta característica, conocida como *truco del núcleo*, permitió extender algunos métodos clásicos de reducción lineal; por ejemplo, *Kernel Principal Components Analysis* (KPCA) [Schölkopf et al., 1998] y *Kernel Sliced Inverse Regression* (KSIR) [Wu, 2008] son extensiones no lineales de PCA y SIR, respectivamente.

Sin embargo, el uso de núcleos reproductores no implica únicamente aplicar un método conocido sobre el espacio característico $\mathcal{H}_\mathcal{X}$. Otra estrategia consiste en caracterizar el

problema de SDR mediante *operadores de covarianza*, los cuales constituyen una extensión funcional del concepto de matrices de covarianza.

Definición 3.15. [Baker, 1973] Sean $\mathbf{X} \in \mathcal{X} \subset \mathbb{R}^p$ y $\mathbf{Y} \in \mathcal{Y} \subset \mathbb{R}^m$ v.a., y sean $\mathcal{H}_\mathcal{X}$ y $\mathcal{H}_\mathcal{Y}$ RKHS con núcleo reproductor $k_\mathcal{X}$ y $k_\mathcal{Y}$, respectivamente. El *operador de covarianza cruzada* $\Sigma_{\mathbf{xy}} : \mathcal{H}_\mathcal{X} \to \mathcal{H}_\mathcal{Y}$ es aquel operador tal que, para todo par de funciones $f \in \mathcal{H}_\mathcal{X}$ y $g \in \mathcal{H}_\mathcal{Y}$, se verifica

$$\langle g, \Sigma_{\mathbf{xy}} f \rangle_{\mathcal{H}_\mathcal{Y}} = \mathbb{E}_{\mathbf{XY}} \left[(f(\mathbf{X}) - \mathbb{E}\left[f(\mathbf{X})\right])(g(\mathbf{Y}) - \mathbb{E}\left[g(\mathbf{Y})\right]) \right].$$

De manera similar, se definen $\Sigma_{\mathbf{yx}}$, $\Sigma_{\mathbf{xx}}$ y $\Sigma_{\mathbf{yy}}$. Además, si $\Sigma_{\mathbf{xx}}$ es invertible, se define el *operador de covarianza condicional* de $\mathbf{Y}$ dado $\mathbf{X}$ mediante

$$\Sigma_{\mathbf{yy}|\mathbf{x}} = \Sigma_{\mathbf{yy}} - \Sigma_{\mathbf{yx}} \Sigma_{\mathbf{xx}}^{-1} \Sigma_{\mathbf{xy}}.$$

Claramente, $\Sigma_{\mathbf{yx}}$ es operador adjunto[2] de $\Sigma_{\mathbf{xy}}$, mientras que $\Sigma_{\mathbf{xx}}$ y $\Sigma_{\mathbf{yy}}$ son autoadjuntos.

Ejemplos de métodos de reducción no lineal que caracterizan los DRS mediante operadores de covarianza son *Manifold Kernel Dimension Reduction* (mKDR) [Nilsson et al., 2007] y *Covariance Operator Inverse Regression* (COIR) [Kim and Pavlovic, 2011]. El primero es una adaptación al campo del *manifold learning* de las ideas de *Kernel Dimension Reduction* (KDR) [Fukumizu et al., 2009], el cual es un método de SDR lineal no paramétrico que se enfoca de forma directa en la independencia condicional de (3.7), la caracteriza en términos del operador de covarianza condicional $\Sigma_{\mathbf{yy}|\mathbf{x}}$ y, a partir de ello, plantea un problema de optimización para estimar el CS. Por su parte, dentro del enfoque de regresión inversa $\mathbf{X}|Y$, COIR permite obtener reducciones no lineales de $\mathbf{X}$ a partir de la caracterización mediante operadores de covarianza de la versión basada en núcleos de $\mathbf{K}_{\text{SIR}}$. Repasaremos este último método en la Subsección 3.3.2.

De los métodos de reducción no lineal mencionados hasta aquí, utilizaremos EF-DR y COIR. A continuación, los describiremos con más detalle.

[2]Es decir, para toda $f \in \mathcal{H}_\mathcal{X}$ y toda $g \in \mathcal{H}_\mathcal{Y}$, se verifica $\langle \Sigma_{\mathbf{yx}} g, f \rangle_{\mathcal{H}_\mathcal{X}} = \langle g, \Sigma_{\mathbf{xy}} f \rangle_{\mathcal{H}_\mathcal{Y}}$.

3.3.1 Reducción suficiente para la familia exponencial (EF-DR) [Bura et al., 2016]

Supongamos que $\mathbf{X}|(Y = y)$ tiene función de densidad en una familia exponencial $\mathcal{P}_{\text{fin}}$ dada por (2.5). Es decir, a cada $y \in \mathcal{Y} \subset \mathbb{R}$ le corresponde $\boldsymbol{\theta}_y \in \boldsymbol{\Theta} \subset \mathbb{R}^m$ tal que

$$\mathbf{X}|(Y = y) \sim p(\mathbf{x}|\boldsymbol{\theta}_y) = \frac{q_0(\mathbf{x})}{Z(\boldsymbol{\theta}_y)} \exp\langle \boldsymbol{\theta}_y, \mathbf{T}(\mathbf{x})\rangle_{\mathbb{R}^m}.$$

Sea $\bar{\boldsymbol{\theta}} = \mathbb{E}_Y[\boldsymbol{\theta}_Y]$. Si $\boldsymbol{\Gamma} \in \mathbb{R}^{m\times d}$ es una matriz semi-ortogonal cuyas columnas forman una base para $\{\boldsymbol{\theta}_y - \bar{\boldsymbol{\theta}} : y \in \mathcal{Y}\}$, podemos escribir

$$\boldsymbol{\theta}_y = \bar{\boldsymbol{\theta}} + \boldsymbol{\Gamma}\boldsymbol{\nu}_y, \tag{3.18}$$

con $\boldsymbol{\nu}_y \in \mathbb{R}^d$ un vector desconocido. Nuevamente se modela $\boldsymbol{\nu}_y$ mediante (3.13), obteniéndose a partir de (3.18) el modelo lineal generalizado de rango reducido

$$\boldsymbol{\theta}_y = \bar{\boldsymbol{\theta}} + \mathbf{D}(\mathbf{f}_y - \mathbb{E}[\mathbf{f}_y]), \tag{3.19}$$

con $\mathbf{D} = \boldsymbol{\Gamma}\boldsymbol{\beta}$, $\boldsymbol{\beta} \in \mathbb{R}^{d\times r}$ y $d \leq \text{mín}\{m, r\}$. El siguiente teorema establece una SDR minimal de $\mathbf{X}$ para la regresión de Y en $\mathbf{X}$, la cual no es lineal en $\mathbf{X}$ pero sí en el estadístico suficiente $\mathbf{T}(\mathbf{X})$ de la familia $\mathcal{P}_{\text{fin}}$. Su demostración está basada en el Teorema de Lehman-Scheffé (Teorema 3.7).

Teorema 3.16. [Bura et al., 2016, Teorema 1] *Para todo $y \in \mathcal{Y}$, supongamos que $\mathbf{X}|(Y = y)$ tiene función de densidad en la familia $\mathcal{P}_{\text{fin}}$ definida en (2.5), con parámetro natural $\boldsymbol{\theta}_y \in \boldsymbol{\Theta}$. Sea $\boldsymbol{\alpha} \in \mathbb{R}^{m\times d}$ tal que $\operatorname{span}\boldsymbol{\alpha} = \operatorname{span}\{\boldsymbol{\theta}_y - \bar{\boldsymbol{\theta}} : y \in \mathcal{Y}\}$. La transformación $\mathbf{R} : \mathbb{R}^p \to \mathbb{R}^d$ definida por*

$$\mathbf{R}(\mathbf{X}) = \boldsymbol{\alpha}^{\mathrm{T}}\mathbf{T}(\mathbf{X})$$

es una SDR minimal de $\mathbf{X}$ para la regresión de Y en $\mathbf{X}$.

Bajo el modelo (3.19) y de acuerdo con el Teorema 3.16, se estima $\mathcal{S}_{Y|\mathbf{X}}$ mediante el MLE $\hat{\boldsymbol{\Gamma}}$ de $\boldsymbol{\Gamma}$, lo cual da lugar al método EF-DR. Este ofrece un par de alternativas, que surgen de hacer suposiciones sobre el rango de $\mathbf{D}$ en (3.19).

Método de estimación	**EF-DR**

Teniendo en cuenta el modelo (3.19):

(1) Si el rango de $\mathbf{D}$ se supone conocido, se obtienen los MLE de $\boldsymbol{\Gamma}$ y $\boldsymbol{\beta}$ mediante el algoritmo IRLS [Yee and Hastie, 2003].

(2) Si el rango de $\mathbf{D}$ se supone desconocido, se obtiene el MLE de $\mathbf{D}$, se estima d a partir de un test asintótico propuesto en [Bura et al., 2016] y finalmente se estiman $\boldsymbol{\Gamma}$ y $\boldsymbol{\beta}$ a partir de la descomposición en valores singulares de $\mathbf{D}$.

Si se opta por la alternativa (1), dado que $d = \dim(\mathcal{S}_{Y|\mathbf{X}})$ es en general desconocida, se estiman $\boldsymbol{\Gamma}$ y $\boldsymbol{\beta}$ para $d = 0, \ldots, \text{mín}\{m, r\}$ y luego se selecciona un modelo mediante algún criterio de selección, como ser el criterio de información Bayesiano (BIC) [Schwarz, 1978] o el criterio de información de Akaike (AIC) [Hurvich and Tsai, 1989].

3.3.2 Covariance Operator Inverse Regression (COIR) [Kim and Pavlovic, 2011]

Supongamos $\mathbf{X} \in \mathcal{X} \subset \mathbb{R}^p$ y $\mathbf{Y} \in \mathcal{Y} \subset \mathbb{R}^m$, y sean $\mathcal{H}_\mathcal{X}$ y $\mathcal{H}_\mathcal{Y}$ RKHS con núcleo reproductor $k_\mathcal{X}$ y $k_\mathcal{Y}$, respectivamente. Además, denotemos $\phi_\mathcal{X}$ y $\phi_\mathcal{Y}$ los mapeos característicos correspondientes y consideremos los operadores de covarianza de la Definición 3.15.

Al igual que KSIR, en COIR se pretende estimar la matriz

$$\mathbf{K}_{\text{COIR}} := \text{var}\left(\mathbb{E}\left[\boldsymbol{\phi}_\mathcal{X}(\mathbf{X})|\mathbf{Y}\right]\right). \tag{3.20}$$

Con la suposición de que $\phi_\mathcal{X}(\mathbf{X})$ está centrado en $\mathcal{H}_\mathcal{X}$, la extensión del Teorema 3.11 nos indica que $\mathbb{E}\left[\phi(\mathbf{X})|Y\right]$ está contenida en el espacio de funciones generado por $\{\Sigma_{\mathbf{xx}} b_i\}_{i=1}^d$, donde $\{b_i\}_{i=1}^d \subset \mathcal{H}_\mathcal{X}$ es una base de funciones que genera el CS, la cual es inducida por el operador $\mathbf{K}_{\text{COIR}}$ y sus autofunciones $\{v_i\}_{i=1}^d \subset \mathcal{H}_\mathcal{X}$ a partir de la relación

$$\Sigma_{\mathbf{xx}} b_i = v_i. \tag{3.21}$$

Bajo leves condiciones, se verifica $\Sigma_{\mathbf{xx}|\mathbf{y}} = \mathbb{E}\left[\mathrm{var}\left(\phi_{\mathcal{X}}(\mathbf{X})|\mathbf{Y}\right)\right]$ [Kim and Pavlovic, 2011, Teorema 3]. Por lo tanto, por ley de varianza total[3], la expresión (3.20) puede reescribirse como

$$\begin{aligned}\mathbf{K}_{\mathrm{COIR}} &= \mathrm{var}\left(\phi_{\mathcal{X}}(\mathbf{X})\right) - \mathbb{E}\left[\mathrm{var}\left(\phi_{\mathcal{X}}(\mathbf{X})|\mathbf{Y}\right)\right] \\ &= \Sigma_{\mathbf{xx}} - \Sigma_{\mathbf{xx}|\mathbf{y}} \\ &= \Sigma_{\mathbf{xy}}\Sigma_{\mathbf{yy}}^{-1}\Sigma_{\mathbf{xy}}. \end{aligned} \tag{3.22}$$

La caracterización (3.22) permite obtener $\hat{\mathbf{K}}_{\mathrm{COIR}}$ a partir de la estimación de los operadores de covarianza, lo cual se hace de manera similar que para el caso de matrices de covarianza muestrales. Por ejemplo, $\hat{\Sigma}_{\mathbf{xy}} = \frac{1}{n}\mathbf{W}_{\mathbf{x}}\mathbf{W}_{\mathbf{y}}^{\mathrm{T}}$, donde

$$\mathbf{W}_{\mathbf{x}} := \left[\phi_{\mathcal{X}}(\mathbf{x}_1), \ldots, \phi_{\mathcal{X}}(\mathbf{x}_n)\right], \qquad \text{y} \qquad \mathbf{W}_{\mathbf{y}} := \left[\phi_{\mathcal{Y}}(\mathbf{y}_1), \ldots, \phi_{\mathcal{Y}}(\mathbf{y}_n)\right].$$

Luego, el objetivo es estimar las direcciones de reducción $b \in \mathcal{H}_{\mathcal{X}}$ y asociarlas a vectores de la forma $\boldsymbol{\beta} \in \mathbb{R}^n$, que permitan proyectar una observación $\mathbf{x} \in \mathbb{R}^p$ de manera más sencilla. Veamos a continuación cómo se obtiene este proceso.

Supongamos que $\mathcal{D}_{\mathbf{x}}$ y $\mathcal{D}_{\mathbf{y}}$ están centralizados en sus espacios característicos[4] y que las autofunciones de $\mathbf{K}_{\mathrm{COIR}}$ son de la forma $v = \sum_{i=1}^{n} \alpha_i \phi_{\mathcal{X}}(\mathbf{x}_i) = \mathbf{W}_{\mathbf{x}}\boldsymbol{\alpha}$, con $\boldsymbol{\alpha} \in \mathbb{R}^n$. Las matrices de Gram, definidas mediante

$$(\mathbf{K}_{\mathbf{x}})_{ij} := k_{\mathcal{X}}(\mathbf{x}_i, \mathbf{x}_j) \qquad \text{y} \qquad (\mathbf{K}_{\mathbf{y}})_{ij} := k_{\mathcal{Y}}(\mathbf{y}_i, \mathbf{y}_j), \tag{3.23}$$

verifican $\mathbf{K}_{\mathbf{x}} = \mathbf{W}_{\mathbf{x}}^{\mathrm{T}}\mathbf{W}_{\mathbf{x}}$ y $\mathbf{K}_{\mathbf{y}} = \mathbf{W}_{\mathbf{y}}^{\mathrm{T}}\mathbf{W}_{\mathbf{y}}$. Además, de (3.22), se deduce

$$\hat{\mathbf{K}}_{\mathrm{COIR}} = \frac{1}{n}\mathbf{W}_{\mathbf{x}}\mathbf{K}_{\mathbf{y}}\left(\mathbf{K}_{\mathbf{y}} + n\varepsilon\mathbf{I}_n\right)^{-1}\mathbf{W}_{\mathbf{x}}^{\mathrm{T}}.$$

Usando estas últimas ecuaciones es fácil ver, premultiplicando por $\mathbf{W}_{\mathbf{x}}^{\mathrm{T}}$ ambos miembros del problema de autofunciones $\hat{\mathbf{K}}_{\mathrm{COIR}}v = \lambda v$, que $\boldsymbol{\alpha}$ es autovector de la matriz

$$\frac{1}{n}\mathbf{K}_{\mathbf{y}}(\mathbf{K}_{\mathbf{y}} + n\varepsilon\mathbf{I}_n)^{-1}\mathbf{K}_{\mathbf{x}}. \tag{3.24}$$

[3]Si $\mathrm{var}\left(X_1\right) < \infty$, entonces $\mathrm{var}\left(X_1\right) = \mathbb{E}\left[\mathrm{var}\left(X_1|X_2\right)\right] + \mathrm{var}\left(\mathbb{E}\left[X_1|X_2\right]\right)$.

[4]La centralización de $\mathbf{K}_{\mathbf{x}}$ (análogamente $\mathbf{K}_{\mathbf{y}}$) es $\tilde{\mathbf{K}}_{\mathbf{x}} = (\mathbf{I}_n - \frac{1}{n}\mathbf{1}_n)\mathbf{K}_{\mathbf{x}}(\mathbf{I}_n - \frac{1}{n}\mathbf{1}_n)$, donde $(\mathbf{1}_n)_{ij} := 1$ [ver Schölkopf et al., 1998, Apéndice B].

Luego de calcular la representación $\boldsymbol{\alpha}$ de la autofunción v, la dirección de reducción $b = \hat{\Sigma}_{\mathbf{xx}}^{-1} v$ se obtiene como sigue: representando $b = \sum_{i=1}^{n} \beta_i \phi_{\mathcal{X}}(\mathbf{x}_i) = \mathbf{W}_{\mathbf{x}} \boldsymbol{\beta}$, con $\boldsymbol{\beta} \in \mathbb{R}^n$, y premultiplicando (3.21) por $\mathbf{W}_{\mathbf{x}}^{\mathrm{T}}$, resulta

$$\boldsymbol{\beta} = n\mathbf{K}_{\mathbf{x}}^{-1} \boldsymbol{\alpha}. \tag{3.25}$$

La clave del método COIR radica en que, para proyectar una nueva observación $\mathbf{x}^*$ sobre una dirección de reducción b, solo se necesitan $\boldsymbol{\beta}$ y $\mathbf{K}_{\mathbf{x}}$. En efecto, tal proyección es

$$\langle b, \phi_{\mathcal{X}}(\mathbf{x}^*) \rangle_{\mathcal{H}_{\mathcal{X}}} = \sum_{i=1}^{n} \beta_i \langle \phi_{\mathcal{X}}(\mathbf{x}_i), \phi_{\mathcal{X}}(\mathbf{x}^*) \rangle_{\mathcal{H}_{\mathcal{X}}} = \sum_{i=1}^{n} \beta_i k_{\mathcal{X}}(\mathbf{x}_i, \mathbf{x}^*).$$

Método de estimación **COIR**

---

1. Calcular las matrices de Gram $\mathbf{K}_{\mathbf{x}}$ y $\mathbf{K}_{\mathbf{y}}$ dadas por (3.23).

2. Para $\varepsilon > 0$, sea $\hat{\mathbf{V}} \in \mathbb{R}^{n \times d}$ la matriz cuyas columnas son los autovectores correspondientes a los d autovalores más grandes de (3.24). De (3.25), el estimador COIR queda representado por la matriz $\hat{\mathbf{B}} \in \mathbb{R}^{n \times d}$ dada por

$$\hat{\mathbf{B}} = n\mathbf{K}_{\mathbf{x}}^{-1} \hat{\mathbf{V}}.$$

3. La reducción de $\mathbf{x} \in \mathbb{R}^p$ es

$$\hat{\mathbf{R}}(\mathbf{x}) = \hat{\mathbf{B}}^{\mathrm{T}} \begin{pmatrix} k_{\mathcal{X}}(\mathbf{x}_1, \mathbf{x}) \\ \vdots \\ k_{\mathcal{X}}(\mathbf{x}_n, \mathbf{x}) \end{pmatrix}.$$

3.4 Otros métodos de reducción de dimensiones

En esta sección haremos una breve reseña sobre un par de métodos de reducción que también utilizaremos durante la etapa experimental en el Capítulo 5. Estos métodos presentan variantes para obtener SDR lineales o no lineales de $\mathbf{X}$ para el problema de regresión de Y en $\mathbf{X}$.

En primer lugar, hacemos mención a un conjunto de técnicas de reducción de dimensiones basadas en *graph embedding*. La idea básica consiste en la construcción de un grafo con información métrica sobre la estructura local de los datos, bajo la suposición de que estos están contenidos en una variedad $\mathcal{M}$ de dimensión mucho menor que p. Ejemplos clásicos son *Isometric Feature Mapping* (ISOMAP) [Tenenbaum et al., 2000] y *Locally Linear Embedding* (LLE) [Roweis and Saul, 2000], los cuales conducen a reducciones no lineales y no supervisadas. Para más información, se recomienda el trabajo [Yan et al., 2007]. Por otra parte, diversos métodos de *graph embedding* fueron unificados en [Shen et al., 2020], dentro de una metodología general que involucra resolver un problema de mínimos cuadrados generalizado. A esta generalización la denominaremos LS-GRAPH.

En segundo lugar, hacemos mención especial a *Principal Support Vector Machines* (PSVM) [Li et al., 2011], un método que combina las ideas de SIR y *Contour Regression* (CR) [Li et al., 2005] con SVM. A continuación, lo describiremos un poco más en detalle.

3.4.1 Principal Support Vector Machines (PSVM) [Li et al., 2011]

La idea básica de PSVM es detectar direcciones a lo largo de las cuales la función de regresión $Y|\mathbf{X}$ no varía (*direcciones de contorno*), basada en el hecho de que dichas direcciones generan el complemento ortogonal de $\mathcal{S}_{Y|\mathbf{X}}$. Los autores proponen identificar las direcciones de contorno a partir de hiperplanos separadores derivados de la aplicación de SVM, ya sea en el espacio original $\mathbb{R}^p$ (SVM lineal) o en un espacio característico $\mathcal{H}_\mathcal{X}$ correspondiente a un núcleo no lineal (SVM no lineal). Luego, estiman $\mathcal{S}_{Y|\mathbf{X}}$ mediante PCA sobre las direcciones normales a los hiperplanos separadores obtenidos.

Ahora bien, para obtener hiperplanos separadores es necesario crear particiones de $\mathcal{Y}$ mediante *slicing*, de modo tal de generar varios problemas de clasificación binaria. Por ejemplo, si se elige un conjunto $\{c_1, \ldots, c_{h-1}\} \subset \mathbb{R}$, a partir de los h problemas de clasificación binarios resultantes de efectuar las particiones $\mathcal{D}_\mathbf{x} = \mathcal{D}_{h_0,1} \cup \mathcal{D}_{h_0,2}$, donde $\mathcal{D}_{h_0,1} := \{\mathbf{x}_i : y_i \leq c_{h_0}\}$ y $\mathcal{D}_{h_0,2} := \{\mathbf{x}_i : y_i > c_{h_0}\}$, se obtienen h direcciones normales.

Es importante remarcar que tanto PSVM lineal como PSVM no lineal están basados en una modificación del problema de optimización asociado a SVM, la cual consiste en añadir

un factor de covarianza a la función objetivo, a efectos de lograr que la implementación de SVM sea invariante respecto a la distribución marginal de $\mathbf{X}$. Como consecuencia de esta modificación, en PSVM no lineal se hace imposible utilizar la versión estándar del algoritmo de SVM. En su lugar, es necesario resolver el problema de optimización modificado a partir de la implementación de algoritmos específicos de programación cuadrática.

3.5 Comentarios de cierre de capítulo

En este capítulo expusimos las ideas y los resultados fundamentales de SDR, completando así el marco conceptual de esta tesis. En particular, repasamos los resultados principales correspondientes al enfoque de regresión inversa basada en modelos, poniendo énfasis en las EF. También revisamos algunos métodos existentes de SDR basados en núcleos, los cuales usaremos más adelante tanto para contrastar los resultados obtenidos con nuestra propuesta como para desarrollar estrategias de reducción en presencia de información adicional.

CAPÍTULO 4

Reducción suficiente de dimensiones para la familia exponencial basada en núcleos

En el Capítulo 2 presentamos en la Definición 2.8 la familia exponencial basada en núcleos (KEF), denotada por $\mathcal{P}$, como una extensión infinito-dimensional de la familia exponencial (EF). Para esta última, en la Subsección 3.3.1 repasamos cómo el enfoque de regresión inversa permitió aprovechar la idea de suficiencia estadística para hallar una SDR de $\mathbf{X}$ para la regresión de Y en $\mathbf{X}$, la cuál no es lineal en $\mathbf{X}$ pero sí en el estadístico suficiente $\mathbf{T}(\mathbf{X})$ de la familia [Bura et al., 2016].

Teniendo en cuenta que, como vimos en la Sección 2.3, las KEF tienen la gran ventaja de ser densas en una clase amplia de distribuciones, es de interés encontrar un resultado de reducción suficiente cuando $\mathbf{X}|Y$ se modela en $\mathcal{P}$. Efectivamente, en la Sección 4.1 obtendremos una SDR de $\mathbf{X}$ para un problema de clasificación $Y|\mathbf{X}$. Trabajaremos dentro del enfoque de regresión inversa y utilizaremos herramientas provenientes de la teoría de estadísticos suficientes, las cuales repasamos en la Sección 3.1.

La reducción suficiente para las KEF no será lineal en el espacio original pero sí en el espacio característico $\mathcal{H}_{\mathcal{X}}$, donde el mapeo característico $\mathbf{x} \mapsto k(\mathbf{x}, \cdot)$ asume un rol fundamental. Una posible desventaja de la SDR que hallaremos es que dependerá del parámetro funcional de la familia, lo cual en la práctica demandaría estimar las densidades condicionales a cada grupo. Afortunadamente, en la Sección 4.2 estableceremos una

conexión con *Máquinas de Vectores Soporte* (SVM) [Boser et al., 1992], la cual será nuestro punto de partida para proponer un método de estimación de la reducción más directo y eficiente, al cual denominaremos RKEF, que prescinde de estimación de densidades. Además, dotará a SVM de propiedades de reducción que no habían sido exploradas.

En las Secciones 4.3 y 4.4 discutiremos la dimensión de la reducción propuesta y los parámetros involucrados, respectivamente, determinando los criterios de selección que utilizaremos en la parte experimental de la tesis. Finalmente, en la Sección 4.5 formalizaremos nuestro método RKEF basado en núcleos.

4.1 Reducción suficiente basada en núcleos

A lo largo de este capítulo, nuestro objeto de estudio será un problema de clasificación $Y|\mathbf{X}$, con $\mathbf{X} \in \mathcal{X} \subset \mathbb{R}^p$ y $Y \in \mathcal{Y} = \{1, \cdots, r\}$, donde $r \in \mathbb{N}$. Asumiremos que, para todo $y \in \mathcal{Y}$, $\pi_y := \mathbb{P}\{Y = y\} \neq 0$ y la distribución condicional de $\mathbf{X}|(Y = y)$ pertenece a la KEF $\mathcal{P}$ dada por (2.10), generada por un espacio de Hilbert $\mathcal{H}_\mathcal{X}$ con núcleo reproductor $k : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}$. Es decir, existe una función $f_y \in \mathcal{H}_\mathcal{X}$ tal que $\mathbf{X}|(Y = y)$ tiene función de densidad

$$p(\mathbf{x}|y) := p_{f_y}(\mathbf{x}) = \frac{q_0(\mathbf{x})}{Z(f_y)} \exp\langle f_y, k(\mathbf{x}, \cdot)\rangle_{\mathcal{H}_\mathcal{X}}. \tag{4.1}$$

El siguiente resultado establece una SDR de $\mathbf{X}$ para la clasificación $Y|\mathbf{X}$. Su demostración se presenta en el Anexo C.1.

Teorema 4.1. *Para todo $y \in \mathcal{Y}$, supongamos que $\mathbf{X}|(Y = y)$ tiene función de densidad $p_{f_y}(\mathbf{x})$ dada por (4.1) en la familia $\mathcal{P}$ definida en (2.10). Para $y \in \mathcal{Y}$ se define la función coordenada*

$$R_y(\mathbf{x}) = (f_y - \mathbb{E}[f])(\mathbf{x}), \tag{4.2}$$

donde $\mathbb{E}[f] := \sum_{i=1}^r \pi_i f_i \in \mathcal{H}_\mathcal{X}$. La transformación $\mathbf{R} : \mathbb{R}^p \to \mathbb{R}^{r-1}$ definida por

$$\mathbf{R}(\mathbf{X}) = (R_1(\mathbf{X}), \ldots, R_{r-1}(\mathbf{X}))^{\mathrm{T}} \tag{4.3}$$

es una SDR de $\mathbf{X}$ para el problema de clasificación $Y|\mathbf{X}$.

De (4.2) y la propiedad reproductora de $\mathcal{H}_\mathcal{X}$ (ver (II) en Definición 2.1 de RKHS), se deduce que cada función coordenada puede escribirse de la forma

$$R_y(\mathbf{x}) = f_y(\mathbf{x}) - \sum_{i=1}^{r} \pi_i f_i(\mathbf{x}) = \sum_{i=1}^{r} \alpha_i \langle f_i, k(\mathbf{x}, \cdot)\rangle_{\mathcal{H}_\mathcal{X}}, \tag{4.4}$$

con coeficientes $a_i = \pi_i$ (excepto para $i = y$, en cuyo caso $\alpha_y = 1 - \pi_y$). La expresión (4.4) nos indica que, en cierto sentido, la reducción $\mathbf{R}(\mathbf{X})$ obtenida en (4.3) es lineal en $k(\mathbf{x}, \cdot)$, quien recordemos asume el rol de estadístico suficiente en la KEF. Dado que, para una EF, el Teorema 3.16 establece una SDR de $\mathbf{X}$ lineal en $\mathbf{T}(\mathbf{X})$, podemos pensar el Teorema 4.1 como una extensión a $\mathcal{P}$ para el caso de problemas de clasificación. Veamos la relación entre los Teoremas 3.16 y 4.1 a modo de ejemplo.

Ejemplo 4.2. Para todo $y \in \mathcal{Y}$, supongamos que $\mathbf{X}|(Y = y)$ tiene función de densidad $p(\mathbf{x}|\boldsymbol{\theta}_y)$ en la familia $\mathcal{P}_{\text{fin}}$ dada por (2.5), con parámetro natural $\boldsymbol{\theta}_y \in \boldsymbol{\Theta}$. Del Ejemplo 2.9, $\mathcal{P}_{\text{fin}}$ puede expresarse como una $\mathcal{P}$ determinada por el RKHS (2.11) de dimensión finita. Así, resulta $p(\mathbf{x}|\boldsymbol{\theta}_y) = p_{f_y} \in \mathcal{P}$ con parámetro funcional $f_y = \langle \boldsymbol{\theta}_y, \mathbf{T}(\cdot)\rangle_{\mathbb{R}^m}$. Además,

$$\mathbb{E}\left[f\right](\mathbf{x}) = \sum_{i=1}^{r} \pi_i \langle \boldsymbol{\theta}_i, \mathbf{T}(\mathbf{x})\rangle_{\mathbb{R}^m} = \langle \bar{\boldsymbol{\theta}}, \mathbf{T}(\mathbf{x})\rangle_{\mathbb{R}^m},$$

con $\bar{\boldsymbol{\theta}} = \mathbb{E}\left[\boldsymbol{\theta}_y\right]$. Por lo tanto, las funciones coordenadas dadas por (4.2) son

$$R_y(\mathbf{x}) = (f_y - \mathbb{E}\left[f\right])(\mathbf{x}) = \langle \boldsymbol{\theta}_y, \mathbf{T}(\mathbf{x})\rangle_{\mathbb{R}^m} - \langle \bar{\boldsymbol{\theta}}, \mathbf{T}(\mathbf{x})\rangle_{\mathbb{R}^m} = \langle \boldsymbol{\theta}_y - \bar{\boldsymbol{\theta}}, \mathbf{T}(\mathbf{x})\rangle_{\mathbb{R}^m}.$$

Entonces, si $\mathbf{A} \in \mathbb{R}^{m\times(r-1)}$ es la matriz cuyas columnas son los $r-1$ vectores $\boldsymbol{\theta}_y - \bar{\boldsymbol{\theta}}$, la SDR de $\mathbf{X}$ para $Y|\mathbf{X}$ dada por (4.3) en el Teorema 4.1 es

$$\mathbf{R}_1(\mathbf{X}) = \mathbf{A}^{\mathrm{T}}\mathbf{T}(\mathbf{X}). \tag{4.5}$$

De (4.5) se puede observar que $\mathbf{R}_1(\mathbf{X})$ es lineal en $\mathbf{T}(\mathbf{X})$, el estadístico suficiente de la familia. Además, si $\boldsymbol{\alpha}$ una base para $\operatorname{span}\{\boldsymbol{\theta}_y - \bar{\boldsymbol{\theta}} : y \in \mathcal{Y}\}$, se verifica

$$\operatorname{span}\mathbf{A} = \operatorname{span}\boldsymbol{\alpha}. \tag{4.6}$$

Ahora bien, del Teorema 3.16 se sabe que $\mathbf{R}_2(\mathbf{X}) = \boldsymbol{\alpha}^{\mathrm{T}}\mathbf{T}(\mathbf{X})$ es una SDR minimal de $\mathbf{X}$ para $Y|\mathbf{X}$. La expresión (4.6) nos asegura que, en $\mathcal{P}_{\text{fin}}$, las SDR que proporcionan los Teoremas 3.16 y 4.1 son equivalentes cuando la matriz $\mathbf{A}$ es de rango completo. □

Un análisis del Teorema 4.1 da lugar a varias observaciones y consecuencias:

Observación 4.3. En la expresión (4.3) de $\mathbf{R}(\mathbf{X})$ se excluye la función coordenada $R_r(\mathbf{X})$ correspondiente al último grupo $y = r$. Sin embargo, es importante remarcar que la idea es *excluir un grupo*, independientemente de cuál sea, ya que la información discriminante que contiene puede recuperarse a partir de los restantes mediante $\mathbb{E}[f]$. Formalizaremos esto con el siguiente resultado, cuya demostración se encuentra en el Anexo C.2.

Proposición 4.4. *Bajo las condiciones del Teorema 4.1, sean* $\mathbf{R}(\mathbf{X})$ *y* $\mathbf{R}'(\mathbf{X})$ *ambas SDR de* $\mathbf{X}$ *para el problema de clasificación* $Y|\mathbf{X}$*, resultantes de excluir grupos diferentes en (4.3). Entonces* $\mathbf{R}(\mathbf{X})$ *y* $\mathbf{R}'(\mathbf{X})$ *son SDR equivalentes.*

Observación 4.5. (Clasificación binaria) Si $r = 2$, la SDR definida en (4.3) es una función real, lo cual significa que es minimal. Formalizamos esto en el siguiente resultado.

Corolario 4.6. *Supongamos* $r = 2$*. Bajo las condiciones del Teorema 4.1, la transformación* $R : \mathbb{R}^p \to \mathbb{R}$ *definida por*

$$R(\mathbf{X}) = (f_1 - f_2)(\mathbf{X}) \tag{4.7}$$

es una SDR minimal de $\mathbf{X}$ *para el problema de clasificación* $Y|\mathbf{X}$*.*

Si bien el Corolario 4.6 es consecuencia directa del Teorema 4.1, se puede demostrar de forma alternativa mediante el Teorema de Lehmann-Scheffé (Teorema 3.7), utilizando un procedimiento similar a la demostración del Teorema 3.16 en [Bura et al., 2016]. Por ese motivo, presentamos dicha prueba alternativa en el Anexo C.3.

Observación 4.7. (Clasificación multiclase 1-vs-1) Si $r > 2$, podemos extender (4.7) tomando los subproblemas binarios 1-vs-1 y definiendo

$$\tilde{\mathbf{R}}(\mathbf{X}) = (f_1 - f_2, \ldots, f_1 - f_r, \ldots, f_{r-1} - f_r)^{\mathrm{T}}(\mathbf{X}). \tag{4.8}$$

Como es de esperar, el siguiente teorema nos asegura que $\tilde{\mathbf{R}}(\mathbf{X})$ también conserva la información de $\mathbf{X}$ sobre Y. Su demostración se presenta en el Anexo C.4.

Teorema 4.8. *Supongamos $r > 2$. Bajo las condiciones del Teorema 4.1, la transformación $\tilde{\mathbf{R}} : \mathbb{R}^p \to \mathbb{R}^{r(r-1)/2}$ definida por (4.8) es una SDR de $\mathbf{X}$ para el problema de clasificación $Y|\mathbf{X}$.*

A pesar de que la SDR del Teorema 4.8 es de mayor dimensión que la del Teorema 4.1, nos permitirá establecer en la próxima sección una conexión con el método de clasificación SVM, y definir a partir de ella una metodología de reducción de dimensiones.

Observar que el Teorema 4.1 requiere estimar cada una de las densidades condicionales $\mathbf{X}|(Y = y) \sim p_{f_y}(\mathbf{x})$ en $\mathcal{P}$, lo cual puede hacerse mediante el método de estimación de [Sriperumbudur et al., 2017] que repasamos en la Subsección 2.3.1. Sin embargo, su aplicación puede derivar en procesos poco prácticos, por lo cual estamos interesados en hallar una estrategia más rápida y eficiente. Nuestra tarea en lo que resta del capítulo será diseñar una metodología adecuada, basada en resultados teóricos de reducción suficiente, que permita explotar el potencial conocido de los clasificadores de vectores soporte.

4.2 Reducción suficiente basada en núcleos vía SVM

Comenzaremos considerando un problema de clasificación binaria $Y|\mathbf{X}$, expresando por conveniencia $Y \in \{-1, +1\}$. Sea $\mathcal{D}_{(\mathbf{x},y)}$ un conjunto de datos de entrenamiento y denotemos con $p_{f^{\pm}}(\mathbf{x})$ la función de densidad de $\mathbf{X}|(Y = \pm 1)$, perteneciente a la familia $\mathcal{P}$ dada por (2.10) y generada por un espacio de Hilbert $\mathcal{H}_{\mathcal{X}}$ con núcleo reproductor k.

El método de clasificación SVM (ver breve resumen en Anexo A.1.1) permite obtener una frontera de decisión, posiblemente no lineal, a partir de un clasificador $\hat{G}(\mathbf{x}) = \text{sign}[\hat{f}(\mathbf{x})]$, donde $\hat{f} : \mathbb{R}^p \to \mathbb{R}$ resuelve el problema de optimización

$$\min_{f \in \mathcal{H}_{\mathcal{X}}} \frac{1}{n} \sum_{i=1}^{n} (1 - y_i f(\mathbf{x}_i))_+ + \lambda \|f\|^2_{\mathcal{H}_{\mathcal{X}}}. \tag{4.9}$$

Podemos generalizar (4.9) definiendo el problema

$$\min_{f\in\mathcal{H}_\mathcal{X}} \frac{1}{n}\sum_{i=1}^{n}[(1-y_i f(\mathbf{x}_i))_+]^q + \lambda\|f\|^2_{\mathcal{H}_\mathcal{X}}, \qquad q>1. \tag{4.10}$$

Observar que si $\hat{f}_q : \mathbb{R}^p \to \mathbb{R}$ es la solución de (4.10), entonces $\hat{f}_q$ tiende asintóticamente, tomando $n\to\infty$, al minimizador de $\mathbb{E}\left[(1-Yf(\mathbf{X}))_+\right]^q$. Si denotamos $\alpha(\mathbf{x}) = \mathbb{P}\{Y=1|\mathbf{X}=\mathbf{x}\}$, el siguiente resultado nos dice que dicho minimizador está asociado a la *regla de Bayes* definida por

$$h^*(\mathbf{x}) := \operatorname{sign}\left[\alpha(\mathbf{x})-\frac{1}{2}\right] = \operatorname{sign}\left[\log\frac{\alpha(\mathbf{x})}{1-\alpha(\mathbf{x})}\right]. \tag{4.11}$$

Lema 4.9. [Lin, 2002, Lema 2.1] *Para todo $q>1$, el minimizador de $\mathbb{E}\left[(1-Yf(\mathbf{X}))_+\right]^q$ es la función $f_q:\mathbb{R}^p\to\mathbb{R}$ definida por*

$$f_q(\mathbf{x}) := \frac{\alpha(\mathbf{x})^{\frac{1}{q-1}} - (1-\alpha(\mathbf{x}))^{\frac{1}{q-1}}}{\alpha(\mathbf{x})^{\frac{1}{q-1}} + (1-\alpha(\mathbf{x}))^{\frac{1}{q-1}}}. \tag{4.12}$$

Además, $\operatorname{sign}[f_q(\mathbf{x})] = \operatorname{sign}[\alpha(\mathbf{x})-1/2]$*, tal que el clasificador* $G_q(\mathbf{x}) := \operatorname{sign}[f_q(\mathbf{x})]$ *es equivalente a la regla de Bayes (4.11).*

Estamos en condiciones de establecer la conexión entre la SDR minimal $R(\mathbf{X})$ dada por (4.7) y el método de clasificación SVM para el caso de un problema de clasificación binaria $Y|\mathbf{X}$. La demostración del siguiente lema se encuentra en el Anexo C.5.

Lema 4.10. *Bajo las condiciones del Teorema 4.1, sea $f^\pm \in \mathcal{H}_\mathcal{X}$ el parámetro natural correspondiente a la distribución condicional de $\mathbf{X}|(Y=\pm1)$. Entonces existe un mapeo uno a uno entre $f_q(\mathbf{X})$ dado por (4.12) y la reducción $R(\mathbf{X}) = (f^+ - f^-)(\mathbf{X})$.*

En virtud del lema anterior, se prueba en el Anexo C.6 el siguiente resultado, que establece que f_q conserva la información acerca de Y contenida en $\mathbf{X}$.

Teorema 4.11. *Bajo las condiciones del Teorema 4.1, la transformación $f_q:\mathbb{R}^p\to\mathbb{R}$ definida por (4.12) es una SDR minimal de $\mathbf{X}$ para el problema de clasificación $Y|\mathbf{X}$.*

Para extender el análisis a un problema multiclase, basta considerar la SDR $\tilde{\mathbf{R}}(\mathbf{X})$ del Teorema 4.8 y aplicar el Lema 4.10 a cada subproblema 1-vs-1, estableciendo así un nuevo resultado de suficiencia. La demostración se presenta en el Anexo C.7.

Teorema 4.12. *Para $q > 1$, sea $\{f_q^{i,j} : \mathbb{R}^p \to \mathbb{R} \,|\, i = 1, ..., r-1; j = i+1, ..., r\}$ el conjunto de minimizadores de $\mathbb{E}\left[(1 - Yf(\mathbf{X}))_+\right]^q$, tal que $f_q^{i,j}$ corresponde al subproblema de clasificación binaria entre las clases i y j. Bajo las condiciones del Teorema 4.1, la transformación $\mathbf{R}_{\text{SVM}} : \mathbb{R}^p \to \mathbb{R}^{r(r-1)/2}$ definida por*

$$\mathbf{R}_{\text{SVM}}(\mathbf{X}) = \left(f_q^{1,2}, ..., f_q^{1,r}, ..., f_q^{r-1,r}\right)^{\mathrm{T}}(\mathbf{X}) \tag{4.13}$$

es una SDR de $\mathbf{X}$ para el problema de clasificación $Y|\mathbf{X}$.

Ahora bien, volviendo al caso $q = 1$ dado por (4.9), la relación entre el minimizador de $\mathbb{E}\left[(1 - Yf(\mathbf{X}))_+\right]$ y la regla de Bayes es algo diferente a la situación cuando $q > 1$.

Lema 4.13. [Lin, 2002, Lema 3.1] *El minimizador de $\mathbb{E}\left[(1 - Yf(\mathbf{X}))_+\right]$ es la regla de Bayes $h^*(\mathbf{x})$ definida en (4.11).*

Sin embargo, en términos de clasificación, este hecho es suficiente para capturar la información predictiva. Esto nos motiva a utilizar la idea del Teorema 4.12 aún en $q = 1$ y definir, mediante SVM tradicional, una estimación de la SDR dada por (4.13).

Definición 4.14. Sea $\left\{\hat{f}_{i,j} : \mathcal{X} \to \mathbb{R} \,|\, i = 1, \ldots, r-1; j = i+1, \ldots, r\right\}$ el conjunto de funciones *score* obtenidas mediante SVM 1-vs-1, a partir de un conjunto de datos $\mathcal{D}_{(\mathbf{x},y)}$. Definimos la *reducción en KEF restringida vía SVM* (RKEF) como

$$\hat{\mathbf{R}}_{\text{SVM}}(\mathbf{X}) = (\hat{f}_{1,2}(\mathbf{X}), \ldots, \hat{f}_{1,r}(\mathbf{X}), \ldots, \hat{f}_{r-1,r}(\mathbf{X}))^{\mathrm{T}}. \tag{4.14}$$

La principal ventaja de $\hat{\mathbf{R}}_{\text{SVM}}$ es que permite el uso de algoritmos eficientes para reducir $\mathbf{X}$, mientras dota a SVM de propiedades de reducción que no han sido exploradas.

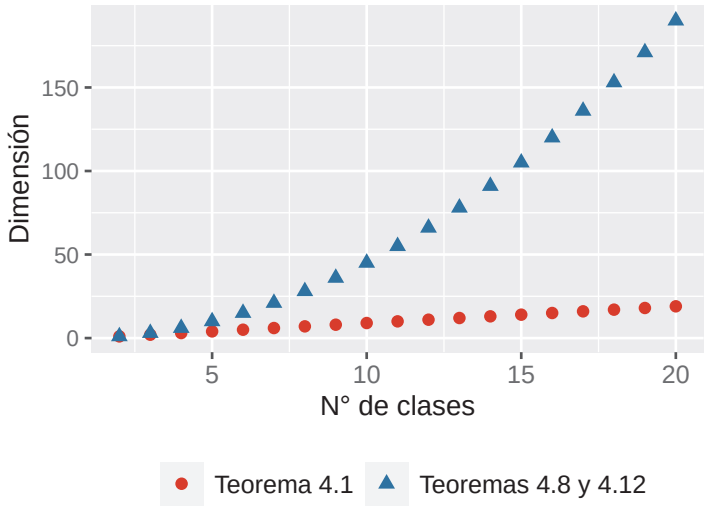

FIGURA 4.1. Dimensión de las SDR obtenidas, en función del N° de clases.

4.3 Dimensión del subespacio de reducción

En la Sección 4.1 presentamos dos SDR de $\mathbf{X}$ para un problema de clasificación $Y|\mathbf{X}$: las reducciones $\mathbf{R}(\mathbf{X})$ y $\tilde{\mathbf{R}}(\mathbf{X})$ de los Teoremas 4.1 y 4.8, respectivamente. Observar que

$$\dim(\tilde{\mathbf{R}}) = \frac{r(r-1)}{2} \geq r - 1 = \dim(\mathbf{R}).$$

Más aún, $\dim(\tilde{\mathbf{R}})$ crece mucho más rápido que $\dim(\mathbf{R})$ en relación a la cantidad r de clases (ver Figura 4.1). Dado que ambas reducciones son suficientes, podemos deducir que en $\tilde{\mathbf{R}}$ hay información redundante. Sin embargo, la diferencia de dimensión entre $\tilde{\mathbf{R}}$ y $\mathbf{R}$ puede ser aliviada utilizando métodos simples de reducción sobre la primera. Por ejemplo, si suponemos que $\tilde{\mathbf{R}}(\mathbf{X})$ se distribuye normalmente, podemos combinarla con PFC [Cook and Forzani, 2008] para obtener reducciones de dimensión menor que $r(r-1)/2$; en particular, $r - 1 = \dim(\mathbf{R})$. Por supuesto, debe quedar claro que en la práctica sugerimos estimar $\tilde{\mathbf{R}}(\mathbf{X})$ mediante la reducción $\hat{\mathbf{R}}_{\text{SVM}}(\mathbf{X})$ dada por (4.14).

Ahora bien, la selección de la dimensión d del DRS puede realizarse a partir de los datos $\mathcal{D}_{(\mathbf{x},y)}$, mediante un proceso de validación cruzada aplicado con algún algoritmo de clasificación sobre las reducciones obtenidas. Más en detalle, el proceso consiste en particionar $\mathcal{D}_{(\mathbf{x},y)}$ en K subconjuntos (K-*fold*), a saber $\mathcal{D}_{(\mathbf{x},y)} = \mathcal{D}_1 \cup \cdots \cup \mathcal{D}_K$, y calcular el *error de validación cruzada*, definido como el promedio de los K errores de clasificación

en $\mathcal{D}_k$ del algoritmo entrenado con $\mathcal{D}_{(\mathbf{x},y)} - \mathcal{D}_k$. Algunos ejemplos de clasificadores clásicos son: Análisis Discriminante Lineal (LDA), Análisis Discriminante Cuadrático (QDA), K vecinos más cercanos (KNN), SVM con núcleo lineal (LSVM) y Perceptrón Multicapa (MLP). Presentamos un breve resumen de ellos en el Anexo A.1.

Es importante remarcar que se espera que las dimensiones

$$d_1 := r(r-1)/2 \qquad \text{y} \qquad d_2 := r - 1, \tag{4.15}$$

derivadas de los Teoremas 4.8 y 4.1, respectivamente, tengan un importante valor en la práctica, en el sentido de que no haya una mejora significativa al analizar otros valores para d. Consideramos que esto daría una importancia extra a los resultados expuestos, por lo cual será una de las cuestiones a analizar durante la parte experimental. De ahora en adelante, con d_1 y d_2 nos referiremos a los valores definidos en (4.15).

4.4 Selección de parámetros

Para obtener la RKEF dada por (4.14), es necesario utilizar un núcleo reproductor para entrenar los clasificadores de SVM. De acuerdo con la teoría expuesta, dicho núcleo será el que genere la familia $\mathcal{P}$ donde se modelan las distribuciones condicionales de $\mathbf{X}|Y$.

Como se puede apreciar en la Tabla 2.1, en general los núcleos reproductores poseen parámetros que deben ser predeterminados. En el caso del núcleo Gaussiano (2.4), para seleccionar el ancho de banda $\sigma \in \mathbb{R}^+$ a partir de $\mathcal{D}_{\mathbf{x}}$, utilizaremos alguno de los siguientes criterios:

- *Criterio 1.* Considerar $\sigma = \sigma_{\text{med}}$, donde σ_{med} es la *mediana heurística* (ver Anexo A.2) definida por

  $$\sigma^2_{\text{med}} := \text{median}\left\{\|\mathbf{x}_i - \mathbf{x}_j\|_2^2 : 1 \leq i < j \leq n\right\}.$$

  Aquí la expresión median M se refiere a la mediana del conjunto de valores M.

- *Criterio 2.* Seleccionar σ de un conjunto finito predeterminado de valores, a partir de un proceso de validación cruzada sobre $\mathcal{D}_{(\mathbf{x},y)}$; es decir, utilizando el error de validación cruzada de algún algoritmo de clasificación.

4.5 Método de reducción RKEF

Sea $\mathcal{D}_{(\mathbf{x},y)}$ un conjunto de datos de entrenamiento correspondiente a un problema de clasificación $Y|\mathbf{X}$. Formalizaremos nuestro método de reducción RKEF, basándonos en los resultados de reducción suficiente de los Teoremas 4.1, 4.8 y 4.12, en la estimación propuesta en la Definición 4.14 y en lo comentado previamente sobre la dimensión d del subespacio de reducción. Para diferenciar de $\hat{\mathbf{R}}_{\text{SVM}}$ en (4.14), denotaremos $\hat{\hat{\mathbf{R}}}_{\text{SVM}}$ cuando la dimensión seleccionada sea menor que d_1.

Método de estimación: **RKEF**

---

1. Obtener $\hat{\mathbf{R}}_{\text{SVM}}(\mathbf{X}) \in \mathbb{R}^{r(r-1)/2}$ de la Definición 4.14. Para SVM Gaussiano, seleccionar $\sigma \in \mathbb{R}^+$ a partir de alguno de los criterios definidos en la Sección 4.4.
2. Elegir $d \leq r(r-1)/2$ mediante un proceso de validación cruzada (ver Sección 4.3). O bien, considerar $d = d_2$ dado por (4.15).
3. Si $d < r(r-1)/2$, aplicar PFC sobre $(\hat{\mathbf{R}}_{\text{SVM}}(\mathbf{X}), Y)$ para obtener $\hat{\hat{\mathbf{R}}}_{\text{SVM}}(\mathbf{X}) \in \mathbb{R}^d$.

Nos referiremos como *reducción vía RKEF* a la reducción obtenida al final del procedimiento, indistintamente si se trata de la reducción inicial $\hat{\mathbf{R}}_{\text{SVM}}$ de dimensión d_1 o de $\hat{\hat{\mathbf{R}}}_{\text{SVM}}$ cuando se seleccione posteriormente $d < d_1$. Para esta última, enfatizamos nuevamente que $d_2 := r - 1$ tiene un importante valor teórico, en virtud del Teorema 4.1.

4.6 Comentarios de cierre de capítulo

En este capítulo desarrollamos un resultado principal de la tesis: un método no lineal de reducción de dimensiones basado en una extensión infinito-dimensional de la familia exponencial. Este método está destinado a preservar la información discriminante presente en los predictores.

Las soluciones típicas de SDR implican transformaciones lineales de los predictores o de formas funcionales bastante rígidas, como es el caso de los estadísticos suficientes de modelos típicos en la EF. Dada la flexibilidad que pueden tener las KEF en función de su núcleo característico, el método presentado representa un aporte significativo en la generalización de los métodos de SDR basados en modelos, permitiendo soluciones altamente no lineales como función de los predictores. Por otra parte, y más allá del interés de este resultado poblacional, mostramos una conexión entre la SDR hallada y el método de clasificación SVM. Esta relación ofrece también una interpretación novedosa de la solución de SVM binario, relacionándolo con el cociente de verosimilitud de modelos en KEF. También tiene importancia práctica ya que es computacionalmente eficiente, admite relaciones p/n grandes y proporciona implícitamente una vía de regularización para limitar el riesgo de sobreajuste a los datos de entrenamiento.

Es justo comentar, de todos modos, que la SDR propuesta garantiza suficiencia pero no minimalidad; es decir, posiblemente exista una reducción menor que aún conserve la información predictiva. En los métodos lineales de SDR, puede abordarse tal problema buscando una factorización de rango reducido de la matriz de proyección inicial. Los métodos basados en KEF carecen de tal matriz y, por lo tanto, la estimación de la reducción minimal es un problema de naturaleza diferente. Como alternativa, sugerimos un enfoque práctico que consiste en post-procesar la RKEF, si se considera necesario, utilizando un método de SDR simple que sí posibilite estimar la dimensión final de forma eficaz. La motivación fundamental para ello es que la reducción de nuestro método es suficiente, con una relación $(r-1)/n$ favorable y con una distribución de las proyecciones resultantes típicamente más fácil de modelar que las variables predictoras originales.

En el próximo capítulo presentaremos una evaluación del método propuesto con datos simulados y con datos reales de alta dimensión.

CAPÍTULO **5**

Simulaciones y ejemplos con datos reales

En este capítulo evaluaremos nuestro método de reducción RKEF en algunas simulaciones y en diferentes conjuntos de datos. A partir de los experimentos iremos explorando las diferentes características emergentes de los capítulos anteriores: RKEF como un procedimiento rápido y eficiente para obtener reducciones cuando se modela $\mathbf{X}|Y$ en la familia exponencial basada en núcleos $\mathcal{P}$, la implementación de RKEF en casos donde la metodología de reducción suficiente para la familia exponencial $\mathcal{P}_{\text{fin}}$ es aplicable y tiene buen rendimiento, y la flexibilidad de $\mathcal{P}$ para aproximar una amplia clase de densidades y modelar diferentes tipos de datos. Esto último nos motiva a aplicar RKEF sobre datos cuya complejidad se deba a su alta dimensionalidad o a alguna otra característica propia de su naturaleza.

Para evaluar nuestro método de reducción y compararlo con otros métodos, utilizaremos distintos algoritmos de clasificación (ver Anexo A.1). Además, también evaluaremos dichos algoritmos en los datos sin reducir, para poder obtener conclusiones respecto a cuán bien las reducciones obtenidas conservan la información predictiva que $\mathbf{X}$ tiene acerca de Y. Debe quedar claro, no obstante, que el interés reside en comparar los métodos de reducción y no en evaluar el desempeño alcanzado por los clasificadores utilizados.

Debido a las buenas propiedades que lo caracterizan, lo cual garantiza no solo la aplicabilidad de RKEF sino de los diferentes métodos basados en núcleos, en todos los casos supondremos que la familia $\mathcal{P}$ está generada por el núcleo Gaussiano (2.4), con ancho

de banda σ seleccionado mediante alguno de los criterios determinados en la Sección 4.4, a los cuales nos referiremos directamente como *Criterio 1* y *Criterio 2* a partir de ahora.

Los experimentos con datos simulados se presentarán en la Sección 5.1, mientras que los experimentos con datos reales conformarán la Sección 5.2. En este último caso se incluyen datos de microbioma (Subsección 5.2.1), datos de cáncer de páncreas (Subsección 5.2.2) y otros conjuntos de datos pertenecientes al ámbito del aprendizaje maquinal (Subsección 5.2.3). La mayoría de estos resultados fueron incluidos en el artículo [Ibañez et al., 2022].

5.1 Simulaciones

Iniciaremos el análisis trabajando con diferentes conjuntos de datos simulados, separados en dos secciones motivadas por distintos propósitos. El primer objetivo es exponer, en la Subsección 5.1.1, la ventaja de estimar la SDR del Teorema 4.1 a partir de RKEF y no mediante la estimación de densidades en $\mathcal{P}$ vía *score matching* (método propuesto en [Sriperumbudur et al., 2017] que detallamos en la Subsección 2.3.1). Luego, nuestro segundo objetivo es comparar el rendimiento de RKEF con algunos métodos clásicos de reducción lineal y no lineal (ver Subsecciones 3.2 y 3.3), utilizando diferentes escenarios de clasificación multiclase.

5.1.1 Ejemplos de clasificación binaria

En esta simulación generamos conjuntos de datos correspondientes a problemas de clasificación binaria. Inicialmente, dichos datos están contenidos en el espacio característico $\mathbb{R}^2$ (ver Figura 5.1), pero son mapeados posteriormente a $\mathbb{R}^{100}$ mediante una matriz aleatoria predeterminada al inicio de la simulación.

En virtud de la Observación 4.5, nuestro objetivo es estimar la reducción unidimensional $R(\mathbf{X})$ de la ecuación (4.7). Lo haremos de dos maneras: (1) $\hat{R}(\mathbf{X})$, resultante de estimar las densidades en $\mathcal{P}$ vía *score matching* mediante el método de [Sriperumbudur et al., 2017] provisto por el Teorema 2.12; y (2) $\hat{R}_{\mathrm{SVM}}(\mathbf{X})$, resultante de aplicar nuestro método RKEF propuesto en el capítulo anterior.

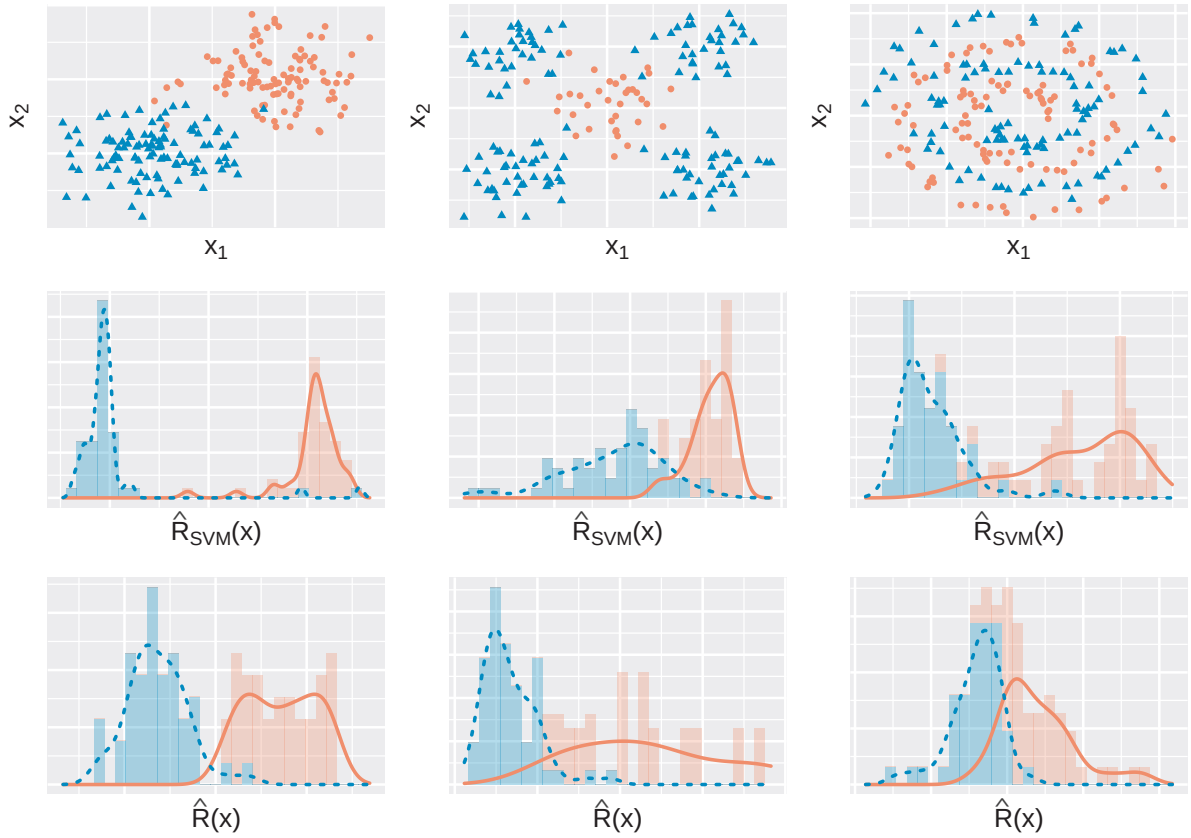

FIGURA 5.1. Ejemplos simulados de clasificación binaria. (Arriba) Datos de entrenamiento generados en $\mathbb{R}^2$. (Medio-Abajo) Distribución de $\hat{R}_{\text{SVM}}(\mathbf{X})$ y $\hat{R}(\mathbf{X})$ en datos de prueba.

Procedimiento *5.1.1. Datos simulados de clasificación binaria*

1° Particionar los datos en 70 % entrenamiento y 30 % prueba.

2° Obtener las reducciones $\hat{R}$ y $\hat{R}_{\text{SVM}}$, ambas de dimensión $\hat{d} = 1$. Utilizar *Criterio 2* para seleccionar el ancho de banda σ del núcleo Gaussiano y la constante de regularización λ en (2.14).

En la Figura 5.1 se muestra la distribución de las reducciones obtenidas para los datos de prueba. Si bien ambas reducciones logran separar las clases, $\hat{R}_{\text{SVM}}$ es más efectiva en el sentido de que las agrupa mejor, logrando una mejor separación entre ellas. Es importante remarcar que esto puede deberse a que la estimación de densidades vía *score matching*

tiene parámetros de regularización difíciles de ajustar en comparación con el método SVM implícito en la reducción vía RKEF. Además, como mencionamos anteriormente, aprovechar la conexión con SVM disminuye significativamente el costo computacional. Por estos motivos, de ahora en adelante elegiremos RKEF como metodología para estimar la SDR cuando las distribuciones condicionales de $\mathbf{X}|Y$ se modelan en una KEF.

5.1.2 Ejemplos de clasificación multiclase

En este experimento consideramos distintos escenarios de clasificación multiclase, determinados a partir de cinco conjuntos de datos bivariados normalmente distribuidos, todos del mismo tamaño. Dichos conjuntos fueron agrupados de diferentes modos en dos o más clases, generando tres escenarios (ver Figura 5.2) con diferentes características:

- *Escenario 1 (SCN1).* Cada conjunto de datos es una clase, resultando en un problema de clasificación de cinco clases. Las densidades condicionales a las clases son normales y, en consecuencia, pertenecen a una familia exponencial $\mathcal{P}_{\text{fin}}$. Para que las clases no sean totalmente separables en $\mathbb{R}^2$, se modificaron levemente los datos para provocar superposición.

- *Escenario 2 (SCN2).* Los datos son agrupados en dos clases que son mezclas de Gaussianas y que no pueden ser separadas por una frontera lineal. En este caso, las densidades condicionales a las clases no pertenecen a una familia $\mathcal{P}_{\text{fin}}$.

- *Escenario 3 (SCN3).* Los datos son agrupados en tres clases, dos de las cuales son mezclas de Gaussianas y la restante tiene distribución normal. En este caso, en una sola clase la densidad condicional pertenece a $\mathcal{P}_{\text{fin}}$.

Posteriormente, los datos son mapeados a $\mathbb{R}^{100}$ mediante una matriz aleatoria predeterminada al inicio del experimento. Dada la forma en que se generan los datos, nos interesa particularmente analizar reducciones unidimensionales. Además, para evaluar el efecto del tamaño de la muestra, se consideraron $\tilde{n}$ observaciones por clase para $\tilde{n} \in \{25, 50, 100\}$. Observar que únicamente en SCN1 las clases son balanceadas.

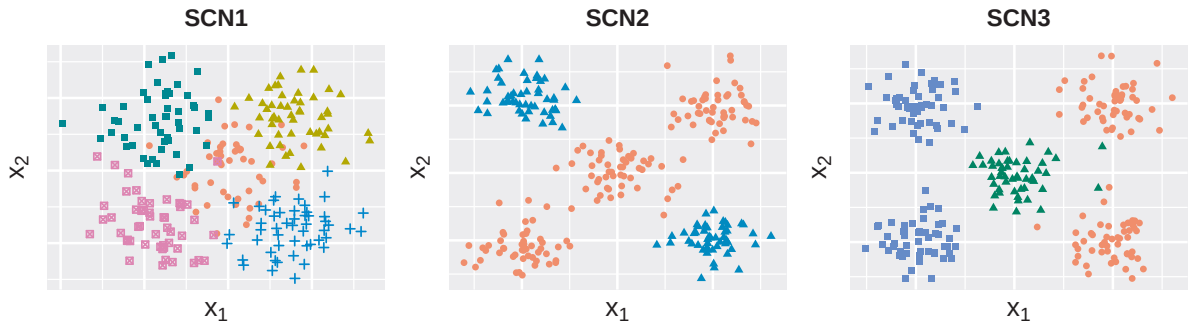

FIGURA 5.2. Escenarios simulados de clasificación multiclase. Los gráficos corresponden al espacio $\mathbb{R}^2$, en el cual se generan los datos.

Procedimiento *5.1.2. Datos simulados de clasificación multiclase*

1° Particionar los datos en 70 % entrenamiento y 30 % prueba.

2° Obtener vía RKEF la reducción $\hat{R}_{\text{SVM}}$, de dimensión $\hat{d} = 1$. Utilizar *Criterio 1* para seleccionar el ancho de banda σ del núcleo Gaussiano.

3° Para comparar con RKEF, obtener reducciones vía COIR (Subsección 3.3.2), SIR (Subsección 3.2.1), PFC (Subsección 3.2.2) y LS-GRAPH.

4° Calcular el error de clasificación en los datos de prueba para un clasificador 5NN.

En la Tabla 5.1 se reportan los resultados obtenidos después de varias repeticiones del experimento. Se observa que:

- En SCN1, los métodos de reducción no lineal RKEF y COIR obtuvieron resultados similares, siendo significativamente mejores a los obtenidos con SIR y PFC, especialmente para el tamaño muestral más pequeño. Sin embargo, SIR y PFC mejoraron su rendimiento para valores más grandes de $\tilde{n}$, como era de esperarse para este tipo de métodos. Es importante remarcar que RKEF mantuvo la capacidad predictiva de los datos, lo cual se deduce de comparar sus resultados con los valores obtenidos por los datos sin reducir.

$\tilde{n}$	Sin reducir	RKEF	COIR	SIR	PFC	LS-GRAPH
SCN1: Cinco clases Gaussianas con densidades en $\mathcal{P}_{\text{fin}}$.						
25	0.225 ± 0.020	$\mathbf{0.219 \pm 0.025}$	0.222 ± 0.021	0.593 ± 0.045	0.590 ± 0.040	0.604 ± 0.043
50	0.203 ± 0.017	0.208 ± 0.017	$\mathbf{0.198 \pm 0.017}$	0.348 ± 0.029	0.344 ± 0.029	0.563 ± 0.035
100	0.199 ± 0.016	0.204 ± 0.020	$\mathbf{0.198 \pm 0.022}$	0.266 ± 0.024	0.264 ± 0.022	0.563 ± 0.047
SCN2: Dos clases mezclas de Gaussianas con densidades en $\mathcal{P}$.						
25	0.101 ± 0.016	0.082 ± 0.020	0.080 ± 0.016	0.360 ± 0.054	0.315 ± 0.057	$\mathbf{0.025 \pm 0.008}$
50	0.091 ± 0.016	0.056 ± 0.009	0.073 ± 0.014	0.162 ± 0.018	0.124 ± 0.020	$\mathbf{0.021 \pm 0.006}$
100	0.093 ± 0.015	0.049 ± 0.011	0.065 ± 0.015	0.102 ± 0.014	0.086 ± 0.012	$\mathbf{0.020 \pm 0.006}$
SCN3: Tres clases con densidades combinadas en $\mathcal{P}_{\text{fin}}$ y $\mathcal{P}$.						
25	0.097 ± 0.020	0.077 ± 0.020	0.077 ± 0.020	0.354 ± 0.042	0.303 ± 0.053	$\mathbf{0.028 \pm 0.008}$
50	0.090 ± 0.013	0.057 ± 0.013	0.077 ± 0.014	0.158 ± 0.018	0.121 ± 0.018	$\mathbf{0.020 \pm 0.008}$
100	0.097 ± 0.014	0.051 ± 0.011	0.064 ± 0.012	0.102 ± 0.012	0.086 ± 0.011	$\mathbf{0.021 \pm 0.007}$

TABLA 5.1. Error de clasificación de 5-NN en escenarios multiclase.

- En SCN2 y en SCN3, nuevamente RKEF y COIR tuvieron mejor rendimiento que los métodos lineales, e inclusive que los datos sin reducir. Más aún, RKEF obtuvo mejores resultados que COIR, lo cual se acentúa más cuando $\tilde{n}$ crece. Sin embargo, el mayor rendimiento es alcanzado por LS-GRAPH, posiblemente debido a una mejor captura de las propiedades locales de los datos.

De acuerdo a las observaciones, concluimos que a pesar de que LS-GRAPH logra captar la información geométrica de SCN2 y SCN3, obteniendo así los mejores resultados del experimento, no consigue evadir el solapamiento de las clases en SCN1. En contrapartida, tanto RKEF como COIR tienen un muy buen rendimiento en todos los escenarios, con una leve ventaja de RKEF cuando el tamaño muestral crece.

Las simulaciones anteriores muestran la flexibilidad de las KEF para modelar diferentes tipos de datos, lo cual se refleja en el buen desempeño del método propuesto RKEF, tanto en los casos donde los métodos clásicos son óptimos como en escenarios de mayor complejidad. En lo que sigue, experimentaremos con datos de problemas reales que involucran situaciones de mayor complejidad.

5.2 Datos reales

5.2.1 Datos de microbioma

En los últimos años hubo un creciente interés en estudiar posibles asociaciones entre la etiología de algunas enfermedades crónicas y la composición de la microbiota que habita el cuerpo humano, determinada por comunidades de microorganismos que conviven en diferentes sectores como la piel, la nariz y la boca. Este entusiasmo ha sido estimulado por nuevas técnicas analíticas de secuenciación genética, que cuantifican de forma precisa la abundancia relativa de los microorganismos que conviven en un entorno determinado.

El Proyecto Microbioma Humano (*Human Microbiome Project*, HMP, `https://www.hmpdacc.org/`), impulsado en 2007 por el Instituto Nacional de Salud (NIH) de Estados Unidos, recolectó durante varios años y en diferentes partes del mundo una gran cantidad de datos de microbioma humano, clasificados según niveles taxonómicos.

Aplicamos RKEF al análisis de un conjunto de datos de HMP, los cuales proporcionan una cuantificación de los microorganismos presentes en diferentes sitios del cuerpo humano: 1) fosas nasales (*anterior nares*), 2) fosa cubital izquierda (*left antecubital fossa*), 3) fórnix posterior (*posterior fornix*), 4) saliva (*saliva*) y 5) heces (*stool*). En este ejemplo, $\mathbf{X}$ es la abundancia relativa de los microorganismos pertenecientes a un nivel taxonómico determinado, mientras que Y es la parte del cuerpo de la cual se extrajo la muestra.

El conjunto de datos $\mathcal{D}_{(\mathbf{x},y)}$ está compuesto por 643 muestras que contienen los resultados correspondientes a dos niveles taxonómicos: filo (*phylum*, nivel L2) y género (*genus*, nivel L6). Los problemas resultantes son de diferente dimensionalidad: $\mathbf{X} \in \mathbb{R}^{22}$ para L2 y $\mathbf{X} \in \mathbb{R}^{551}$ para L6.

5.2.1.1. Nivel taxonómico filo (L2)

Los datos de microbioma a nivel filo (L2) comprenden 22 taxones ($p = 22$), por lo cual estamos ante un problema de baja dimensionalidad. Sin embargo, tienen la dificultad de ser datos composicionales (es decir, proporciones calculadas a partir de vectores

de conteo) y además poseen una elevada proporción de valores nulos (81.99 %, específicamente). Estas características suelen afectar el uso de métodos clásicos. En particular, estos datos no pueden ser modelados mediante una EF y, en consecuencia, no es un caso que esté cubierto por el método EF-DR de [Bura et al., 2016]. Cabe mencionar que en [Tomassi et al., 2019] proponen un modelo específico para datos composicionales, que si bien extiende el enfoque de [Bura et al., 2016] para ese caso, no logra lidiar con la alta dimensionalidad del nivel L6, razón por la cual decidimos no utilizarlo.

La alta proporción de ceros en la matriz de datos imposibilita poder aplicar el método EF-DR. Por ese motivo, precondicionamos el problema premultiplicando los datos por una matriz $\mathbf{A} \in \mathbb{R}^{p\times p}$ aleatoria, predeterminada al inicio del experimento. Este precondicionamiento acerca la distribución de los datos a la normalidad y ayuda a mitigar el problema de exceso de ceros. Bajo estas nuevas condiciones, el método EF-DR es aplicable y la comparación con RKEF se hace factible, lo cual nos permite analizar con este ejemplo cómo se desempeña la generalización infinito-dimensional $\mathcal{P}$ de $\mathcal{P}_{\text{fin}}$ en términos de los resultados de SDR propios de cada caso.

Procedimiento *5.2.1.1. Datos de microbioma a nivel filo (L2)*

1° Particionar los datos para efectuar validación cruzada 10-*fold*.

2° Obtener vía RKEF la reducción $\hat{\mathbf{R}}_{\text{SVM}}$, de dimensión $d_1 = 10$. Utilizar *Criterio 2* para seleccionar el ancho de banda σ del núcleo Gaussiano.

3° Adicionalmente, obtener vía RKEF las reducciones $\hat{\hat{\mathbf{R}}}_{\text{SVM}}$, de dimensiones $\hat{d} \leq 9$. En particular, $d_2 = 4$.

4° Para comparar con RKEF, obtener reducciones vía EF-DR (Subsección 3.3.1).

5° Calcular el error de clasificación 10-*fold* CV para los clasificadores LDA, QDA y LSVM. Utilizar *Criterio 2* para seleccionar el parámetro de costo en LSVM. Luego, determinar en cada caso la dimensión óptima $\hat{d}_{\text{opt}}$.

	$p = 22$	$d_1 = 10$		$d_2 = 4$	
Clasif.	Sin reducir	RKEF	EF-DR	RKEF	EF-DR
LDA	0.160 ± 0.035	$\mathbf{0.135 \pm 0.030}$	0.156 ± 0.031	$\mathbf{0.140 \pm 0.034}$	0.156 ± 0.031
QDA	0.548 ± 0.051	$\mathbf{0.145 \pm 0.039}$	0.215 ± 0.080	0.139 ± 0.034	$\mathbf{0.134 \pm 0.028}$
LSVM	0.121 ± 0.047	0.139 ± 0.035	$\mathbf{0.115 \pm 0.027}$	0.137 ± 0.031	$\mathbf{0.121 \pm 0.030}$

Tabla 5.2. Error de clasificación en datos de microbioma a nivel filo (L2).

	RKEF		EF-DR	
Clasificador	$\hat{d}_{\text{opt}}$	Error	$\hat{d}_{\text{opt}}$	Error
LDA	9	$\mathbf{0.131 \pm 0.032}$	6	0.153 ± 0.036
QDA	4	0.139 ± 0.034	3	$\mathbf{0.126 \pm 0.032}$
LSVM	8	0.128 ± 0.027	8	$\mathbf{0.107 \pm 0.033}$

Tabla 5.3. Dimensión óptima en datos de microbioma a nivel filo (L2).

En la Tabla 5.2 se reportan los resultados obtenidos para $d_1 = 10$ y $d_2 = 4$, mientras que para la dimensión óptima $\hat{d}_{\text{opt}}$ se muestran en la Tabla 5.3. Se observa que:

- En LDA, ambos métodos de reducción logran mejorar el rendimiento respecto a los datos sin reducir. Además, mantienen el poder predictivo desde la dimensión $d_1 = 10$ a la dimensión $d_2 = 4$. El mejor resultado lo obtiene RKEF, con una mejora relativa en media de $(0.153 - 0.131)/0.153 = 14.4\,\%$ respecto de EF-DR.
- En QDA, ambos métodos de reducción logran solucionar los problemas del clasificador QDA en los datos sin reducir. El mejor resultado lo logra EF-DR, con una mejora relativa en media de $(0.139 - 0.126)/0.139 = 9.4\,\%$. Sin embargo, se observa que RKEF conserva la estabilidad de QDA de $d_1 = 10$ a $d_2 = 4$, mientras que EF-DR obtiene en $d_1 = 10$ el error de clasificación más alto (0.215 ± 0.080).
- En LSVM, otra vez los métodos conservan el rendimiento en los datos sin reducir. Hay una leve ventaja en favor de EF-DR, que en su mejor resultado obtiene una mejora relativa en media de $(0.128 - 0.107)/0.128 = 16.4\,\%$ respecto a RKEF.

De acuerdo a las observaciones, concluimos que ambos métodos de reducción logran retener la información predictiva de $\mathbf{X}$ acerca de Y, alcanzando buenos rendimientos tanto en $d_1 = 10$ como en $d_2 = 4$. Además, ambos solucionan el problema de QDA sobre los datos sin reducir, pasando de un muy mal error medio de 0.548 a excelentes resultados (< 0.140 en ambos casos) en dimensión $d_2 = 4$.

Por otra parte, vemos que EF-DR logra levemente una mejor compresión de los datos, logrando dimensiones óptimas menores para la reducción. De todos modos, las tasas de error obtenidas con la dimensión óptima estimada son similares a las reportadas con d_1 o d_2. Mas aún, resultó $\hat{d}_{\text{opt}} > d_2$, excepto en un solo caso. Por esta razón, en el futuro reportaremos únicamente los resultados para d_1 y d_2.

Dado que el escenario que presentan los datos de microbioma L2 es propicio para las suposiciones de EF-DR, podemos afirmar en favor de nuestro método propuesto RKEF que logra ser comparable con el método clásico en situaciones de baja dimensionalidad, lo cual era deseable teniendo en cuenta que presentamos $\mathcal{P}$ como una extensión de $\mathcal{P}_{\text{fin}}$.

A continuación, en el experimento para el nivel taxonómico L6, abordamos una extensión de mayor interés práctico en la que el tamaño muestral es el mismo, pero tanto la dimensionalidad como la proporción de ceros aumenta significativamente.

5.2.1.2. Nivel taxonómico género (L6)

Los datos de microbioma a nivel género (L6) comprenden 551 taxones ($p = 551$). En este caso, las técnicas de reducción basadas en modelos de la familia exponencial $\mathcal{P}_{\text{fin}}$ se ven fuertemente afectadas por la dimensión de los datos. En consecuencia, aún transformándolos para intentar esquivar la estructura composicional y la alta esparsidad, el método EF-DR no es aplicable para este nivel taxonómico. Más aún, al pasar del nivel filo (L2) al nivel género (L6), aumenta la proporción de ceros (de 81.99 % a 92.51 %), lo cual desafía aún más la capacidad de los modelos multivariados que no son adaptados a modelos inflados en cero. Por lo tanto, precondicionaremos el problema de la misma forma que para el nivel taxonómico L2.

	Sin reducir ($p = 551$)				
	LDA: 0.361 ± 0.054	LSVM: 0.044 ± 0.022		MLP: 0.042 ± 0.022	
	$d_1 = 10$				
Clasificador	RKEF	COIR	PLS	PSVM	LS-GRAPH
LDA	$\mathbf{0.047 \pm 0.020}$	0.051 ± 0.024	0.054 ± 0.027	0.089 ± 0.030	0.067 ± 0.024
LSVM	0.044 ± 0.025	0.044 ± 0.020	$\mathbf{0.040 \pm 0.022}$	0.065 ± 0.025	0.053 ± 0.025
MLP	0.045 ± 0.021	0.048 ± 0.026	$\mathbf{0.042 \pm 0.020}$	0.082 ± 0.034	0.053 ± 0.024
	$d_2 = 4$				
Clasificador	RKEF	COIR	PLS	PSVM	LS-GRAPH
LDA	$\mathbf{0.047 \pm 0.020}$	0.072 ± 0.017	0.115 ± 0.018	0.137 ± 0.030	0.143 ± 0.025
LSVM	$\mathbf{0.041 \pm 0.027}$	0.056 ± 0.024	0.072 ± 0.017	0.114 ± 0.035	0.098 ± 0.016
MLP	$\mathbf{0.045 \pm 0.019}$	0.058 ± 0.015	0.067 ± 0.021	0.114 ± 0.035	0.079 ± 0.033

TABLA 5.4. Error de clasificación en datos de microbioma a nivel género (L6).

Procedimiento *5.2.1.2. Datos de microbioma a nivel género (L6)*

1° Particionar los datos para efectuar validación cruzada 10-*fold*.

2° Obtener vía RKEF las reducciones $\hat{\mathbf{R}}_{\text{SVM}}$, de dimensión $d_1 = 10$, y $\hat{\hat{\mathbf{R}}}_{\text{SVM}}$, de dimensión $d_2 = 4$. Utilizar *Criterio 2* para seleccionar el ancho de banda σ del núcleo Gaussiano.

3° Para comparar con RKEF, obtener reducciones vía COIR, PLS, PSVM (Subsección 3.4.1) y LS-GRAPH.

4° Calcular el error de clasificación 10-*fold* CV para los clasificadores LDA, LSVM y MLP. Utilizar *Criterio 2* para seleccionar el parámetro de costo en LSVM y el número de neuronas en la capa oculta de MLP.

En la Tabla 5.4 y la Figura 5.3 se reportan los resultados obtenidos. Se observa que:

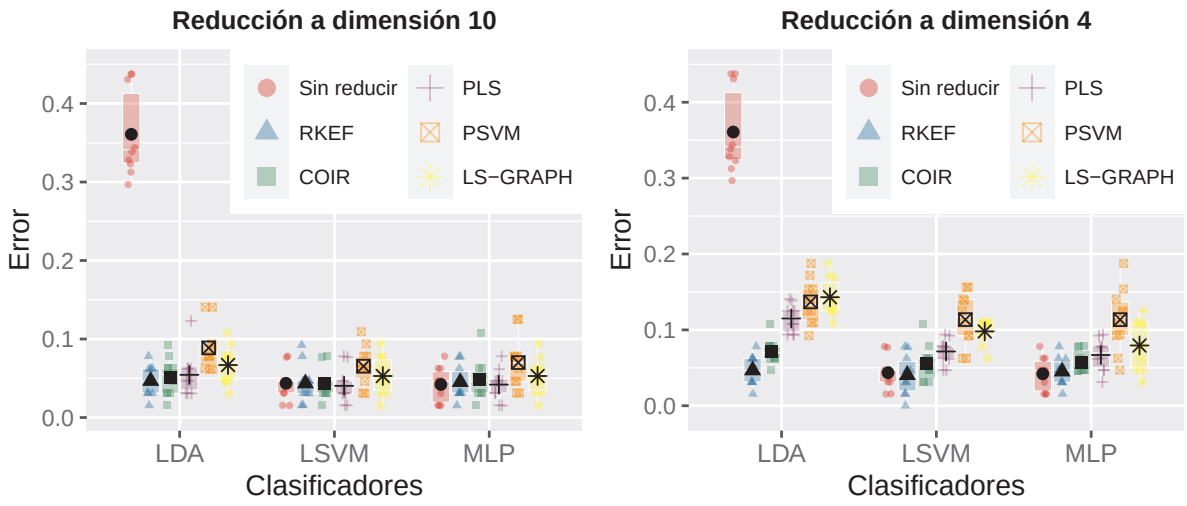

FIGURA 5.3. Error de clasificación en datos de microbioma a nivel género (L6).

- En LDA, los métodos RKEF, COIR y PLS tienen buen rendimiento al reducir a dimensión $d_1 = 10$, mejorando significativamente la eficacia de LDA respecto a cuando fue aplicado sobre los datos sin reducir, con una leve ventaja en favor de RKEF. Por otra parte, en las reducciones a dimensión $d_2 = 4$, nuestro método RKEF logra mantener el buen rendimiento de d_1, con una mejora relativa de $(0.072 - 0.047)/0.072 = 34.7\,\%$ respecto al segundo mejor método (COIR). El resto de los métodos empeoró significativamente su desempeño en d_2.

- En LSVM y MLP la situación es similar a LDA, salvo que ambos clasificadores obtienen muy buenos resultados en los datos sin reducir. En estos casos, RKEF logra mantener el poder predictivo en $d_1 = 10$ y $d_2 = 4$. Además, aunque PLS tiene una leve ventaja a su favor en d_1, nuevamente en d_2 disminuye su eficacia. Por su parte, PSVM y LS-GRAPH siguen teniendo los peores resultados, mientras que COIR y RKEF son los que menos sufren el cambio de dimensión de d_1 a d_2.

De acuerdo a las observaciones, concluimos que tanto RKEF como COIR logran preservar la capacidad predictiva de $d_1 = 10$ a $d_2 = 4$, con una leve ventaja en favor de nuestro método. Podemos afirmar que este ejemplo es una evidencia del potencial de los métodos basados en núcleos. En cuanto a RKEF, hay una clara influencia del Teorema 4.1, que sugiere redundancia en $d_1 = 10$, eliminada al aplicar PFC para alcanzar $d_2 = 4$.

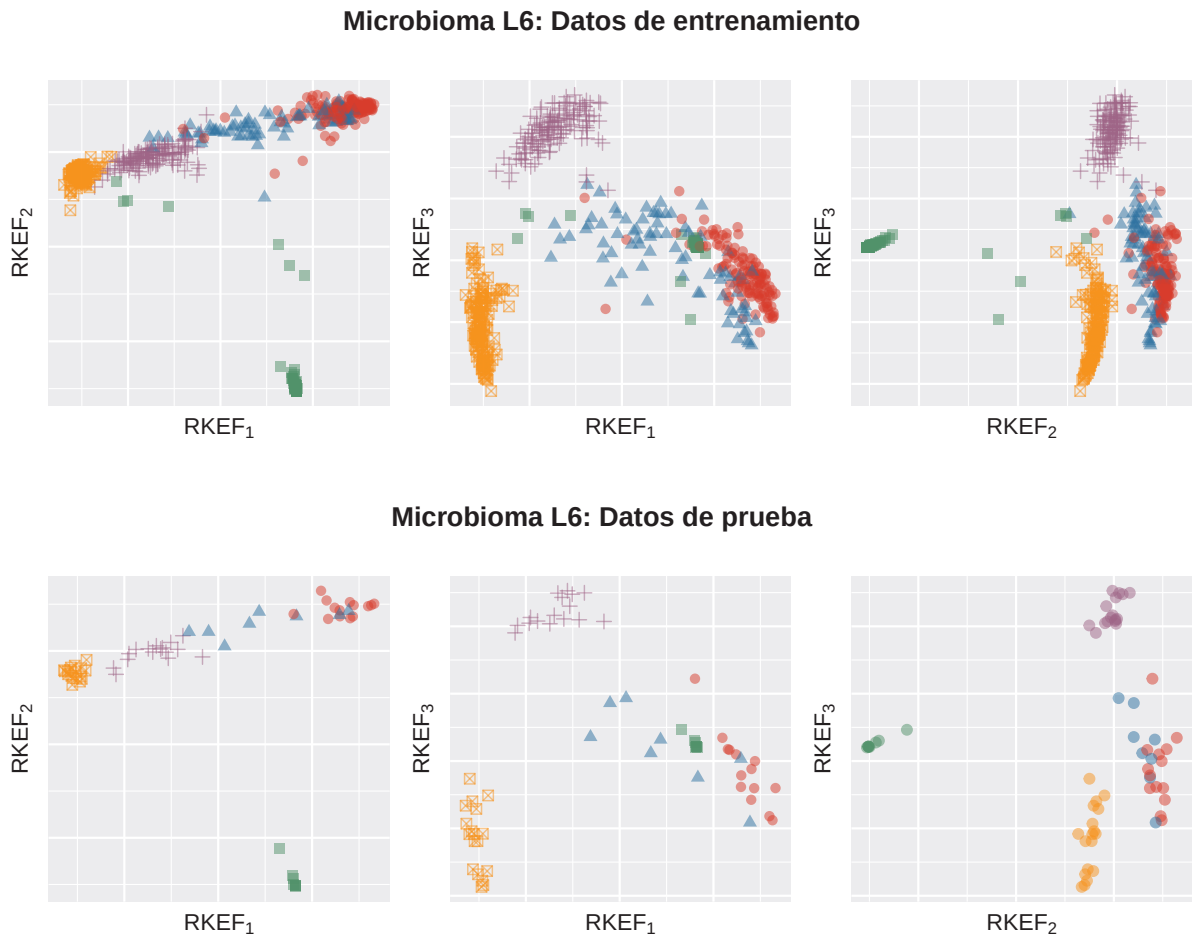

FIGURA 5.4. Gráficas por pares de las tres primeras PFC de $\hat{\mathbf{R}}_{\text{SVM}}$ en datos de microbioma a nivel género (L6).

Por otra parte, destacamos también la favorable comparación directa de nuestro método con PSVM [Li et al., 2011], especialmente en $d_2 = 4$, donde RKEF logra mejoras significativas respecto a PSVM: $(0.137-0.047)/0.137 = 65.7\,\%$ en LDA, $(0.114-0.041)/0.114 = 64\,\%$ en LSVM y $(0.114 - 0.045)/0.114 = 60.5\,\%$ en MLP. Esta comparación tiene valor si se tiene en cuenta que ambos métodos están basados en SVM.

Por último, para ilustrar que la reducción $\hat{\mathbf{R}}_{\text{SVM}}$ no sufre de *overfitting*, incluimos en la Figura 5.4 la gráfica de sus primeras tres PFC, tanto para los datos de entrenamiento como para lo datos de prueba, correspondientes a una de las particiones utilizadas en el experimento. Allí podemos observar que las características geométricas de la reducción son similares en datos de prueba y entrenamiento, y también visualizar la concentración de las clases, lo cual facilita la tarea posterior de los algoritmos de clasificación.

5.2.2 Datos de cáncer de páncreas

El Consorcio Internacional del Genoma del Cáncer (*International Cancer Genome Consortium*, ICGC, `https://dcc.icgc.org`) es una organización científica creada en el año 2008, cuyo principal objetivo es estudiar los genomas de cánceres de interés mundial. Las diferentes naciones participantes financian el estudio de al menos 50 tipos y/o subtipos de cáncer de carácter clínico, los cuales comprenden aproximadamente 25000 genomas de cáncer. El propósito es optimizar el pronóstico y manejo terapéutico del cáncer, y promover el desarrollo de nuevas terapias. Los proyectos del ICGC analizan tipos de tumores que afectan diferentes partes del cuerpo, como la sangre, el cerebro, las mamas, los riñones, el hígado, el páncreas, el estómago, la cavidad oral y los ovarios [ICGC, 2010].

Analizamos un conjunto de datos correspondientes a cáncer de páncreas, al cual nombraremos PanCancer. Consiste en 4333 muestras de entrenamiento y 1089 muestras de prueba, correspondientes a mediciones de más de 19000 variables predictoras que contienen información acerca de 10 tipos de cáncer de páncreas. Es decir, estudiamos un problema de clasificación con $Y \in \{1, \ldots, 10\}$. Para condicionar el problema, premultiplicamos la matriz de datos por una matriz $\mathbf{A} \in \mathbb{R}^{1000\times p}$, lo cual conduce a $\mathbf{X} \in \mathbb{R}^{1000}$. Aún con esta primera reducción aleatoria, seguimos en un escenario de alta dimensionalidad.

Dado que en el ejemplo de datos de microbioma L6 de la sección anterior, los métodos RKEF, COIR y PLS tuvieron rendimientos similares, especialmente en $d_1 = 10$, nuestro propósito es realizar una nueva comparación entre ellos. De hecho, ahora la dimensión ($p = 1000$) será más exigente, particularmente para PLS, más aún considerando el rango de dimensiones que impondrá RKEF.

Procedimiento *5.2.2. Datos PanCancer*

1° Obtener vía RKEF las reducciones $\hat{\mathbf{R}}_{\text{SVM}}$, de dimensón $d_1 = 45$, y $\hat{\mathbf{R}}_{\text{SVM}}$, de dimensión $d_2 = 9$. Utilizar *Criterio 1* para seleccionar el ancho de banda σ del núcleo Gaussiano.

2° Para comparar con RKEF, obtener reducciones vía COIR y PLS.

Sin reducir ($p = 1000$)			
LDA: 0.087 ± 0.004	QDA: 0.692 ± 0.008		LSVM: 0.037 ± 0.022
	$d_1 = 45$		
Clasificador	RKEF	COIR	PLS
LDA	$\mathbf{0.070 \pm 0.003}$	0.110 ± 0.014	0.082 ± 0.009
QDA	$\mathbf{0.138 \pm 0.013}$	0.143 ± 0.012	0.136 ± 0.027
LSVM	0.054 ± 0.005	$\mathbf{0.049 \pm 0.006}$	0.053 ± 0.016
	$d_2 = 9$		
Clasificador	RKEF	COIR	PLS
LDA	$\mathbf{0.070 \pm 0.003}$	0.137 ± 0.005	0.228 ± 0.016
QDA	$\mathbf{0.101 \pm 0.003}$	0.107 ± 0.005	0.202 ± 0.023
LSVM	0.095 ± 0.021	$\mathbf{0.091 \pm 0.011}$	0.119 ± 0.038

TABLA 5.5. Error de clasificación en datos PanCancer.

3° Calcular el error de clasificación en los datos de prueba de los clasificadores LDA, QDA y LSVM. Utilizar *Criterio 2* para seleccionar el parámetro de costo en LSVM.

En la Tabla 5.5 se reportan los resultados obtenidos luego de varias iteraciones del experimento, donde al inicio de cada una se generó la matriz aleatoria $\mathbf{A}$. Se observa que:

- En LDA, los tres métodos obtienen buenos resultados, pero solo RKEF mantiene el poder predictivo cuando se reduce a $d_2 = 9$. Allí, nuestro método logra una mejora relativa en media de $(0.137 - 0.070)/0.137 = 48.9\,\%$ respecto a COIR.
- QDA tiene graves problemas al ser aplicado sobre los datos sin reducir. Aunque los métodos de reducción aplicados lograron corregir esta situación, no obtienen buenos resultados en comparación al resto de los clasificadores. Los mejores rendimientos son de RKEF y COIR, con una leve ventaja en favor de nuestro método, que en $d_2 = 9$ mejora un $(0.107 - 0.101)/0.107 = 5.6\,\%$ el resultado de COIR.

- LSVM logra excelentes resultados en los datos sin reducir, manteniendo el poder predictivo para $d_1 = 45$ en los tres métodos de reducción. Pero PLS tiene dificultades cuando se trabaja en $d_2 = 9$, por lo cual nuevamente RKEF y COIR logran los mejores resultados. En este caso, COIR logra en $d_2 = 9$ una mejora de $(0.095 - 0.091)/0.095 = 4.2\,\%$ respecto de nuestro método.

De acuerdo a las observaciones, concluimos que si bien no hay grandes diferencias en $d_1 = 45$, podemos destacar el rendimiento de RKEF en $d_2 = 9$, teniendo en cuenta que hay un amplio margen entre ambas dimensiones. Por otra parte, el método PLS resulta ser mucho más sensible al cambio en la dimensión de la reducción, de igual manera a lo que sucedió con los datos de microbioma. Mientras tanto, COIR demuestra su potencial logrando un rendimiento similar a RKEF y, salvo en LDA, tampoco se ve afectado por el cambio de dimensión de $d_1 = 45$ a $d_2 = 9$.

Podemos decir que los métodos RKEF y COIR, que en su algoritmo logran captar las similaridades de los datos a través de los núcleos reproductores, son menos sensibles a la dimensionalidad de los datos. En favor de RKEF, enfatizamos que tiene menos parámetros y se basa en una metodología más intuitiva y sencilla de implementar.

5.2.3 Otros datos

A continuación, presentamos la evaluación utilizando diversos conjuntos de datos (resumidos en la Tabla 5.6): *Liver Disorders*, *Ionosphere* y *Satlog* del repositorio de la Universidad de California (UCI), y *Birds-Planes-Cars* de [Cook and Forzani, 2009]. Con base en los resultados de los ejemplos anteriores, realizamos únicamente reducciones a dimensión $d_2 = r - 1$.

Procedimiento *5.2.3. Otros datos*

1° Particionar los datos para efectuar validación cruzada 10-*fold*.

2° Obtener vía RKEF la reducción $\hat{\hat{\mathbf{R}}}_{\mathrm{SVM}}$, de dimensión $d_2 = r - 1$. Utilizar *Criterio 1* para seleccionar el ancho de banda σ del núcleo Gaussiano.

Datos	n	p	r
Liver Disorders	345	6	2
Ionosphere	351	32	2
Satlog	4435	36	6
Birds-Planes-Cars	165	13	3

TABLA 5.6. Descripción de datos de Subsección 5.2.3.

Datos	Sin reducir	RKEF	COIR	LS-GRAPH
Liver Disorders	0.384 ± 0.076	$\mathbf{0.344 \pm 0.096}$	0.394 ± 0.059	0.439 ± 0.126
Ionosphere	0.165 ± 0.061	$\mathbf{0.054 \pm 0.025}$	0.238 ± 0.126	0.241 ± 0.063
Satlog	0.091 ± 0.014	$\mathbf{0.089 \pm 0.011}$	0.164 ± 0.017	0.113 ± 0.014
Birds-Planes-Cars	0.067 ± 0.054	0.036 ± 0.041	$\mathbf{0.018 \pm 0.029}$	0.120 ± 0.092

TABLA 5.7. Error de clasificación de 5NN en varios conjuntos de datos.

3° Para comparar con RKEF, obtener reducciones vía COIR y LS-GRAPH.

4° Calcular el error de clasificación 10-*fold* CV de un clasificador de 5NN.

En la Tabla 5.7 y la Figura 5.5 se reportan los resultados obtenidos. Se observa que:

- En *Liver Disorders*, las reducciones son de dimensión $d_2 = 1$. Nuestro método alcanza el mejor resultado, seguido de COIR, respecto al cual RKEF logra una mejora relativa en media de $(0.394 - 0.344)/0.394 = 12.7\,\%$.
- En *Ionosphere*, las reducciones también son de dimensión $d_2 = 1$, y en esta situación la diferencia con la dimensión de los datos sin reducir es mayor ($p = 32$). Nuevamente RKEF alcanza el mejor resultado, pero ahora significativamente por encima del resto, con una mejora relativa sobre el segundo mejor método, COIR, de $(0.164 - 0.054)/0.164 = 67.1\,\%$.
- En *Satlog*, las reducciones son de dimensión $d_2 = 5$. Nuestro método RKEF logra mantener la efectividad predictiva respecto de los datos sin reducir y vuelve a

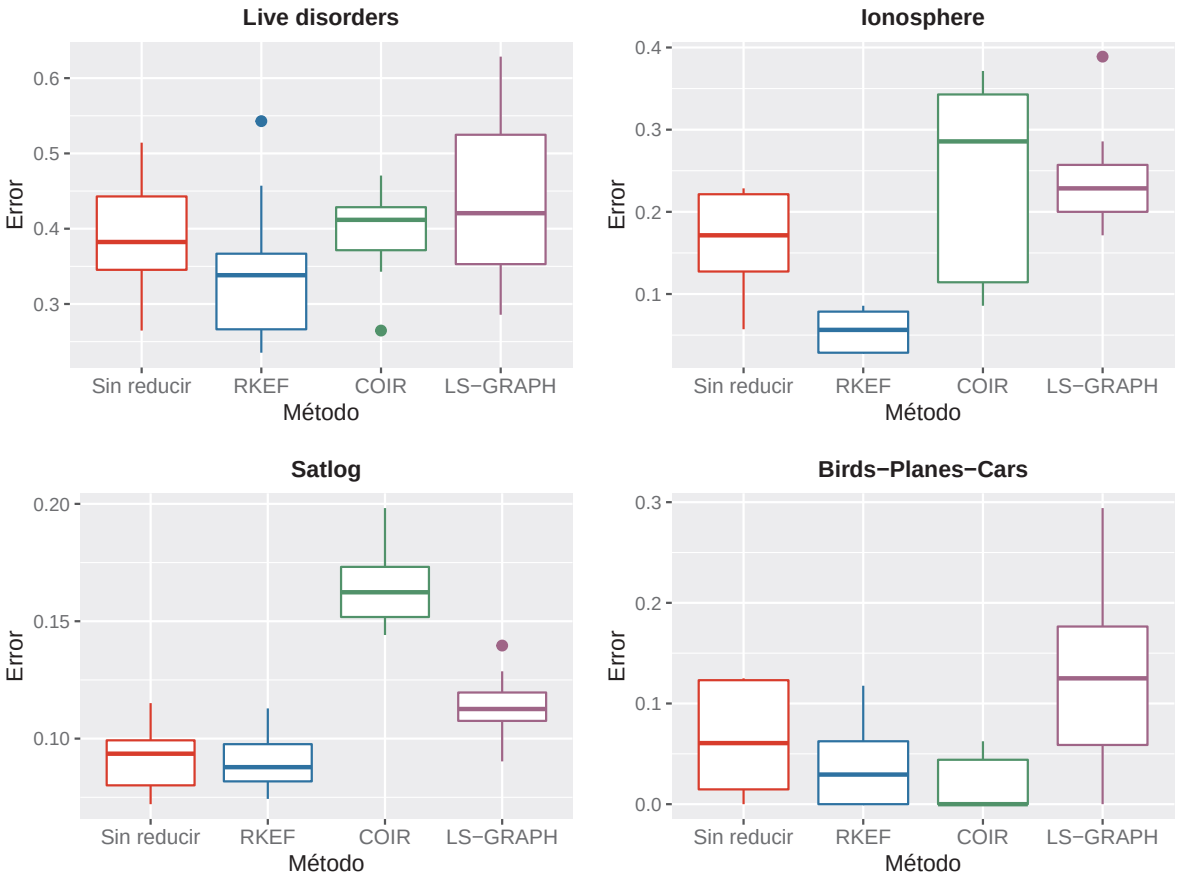

FIGURA 5.5. Error de clasificación de 5NN en varios conjuntos de datos.

obtener el mejor rendimiento, seguido por LS-GRAPH, respecto al cual alcanza una mejora relativa de $(0.113 - 0.089)/0.113 = 21.2\,\%$.

- En *Birds-Planes-Cars*, las reducciones son de dimensión $d_2 = 2$. En este caso, COIR obtiene el mejor resultado, seguido de RKEF, respecto al cual tiene una mejora relativa de $(0.036 - 0.018)/0.036 = 50\,\%$. Por su parte, LS-GRAPH tiene el peor resultado por amplia diferencia.

De acuerdo a las observaciones, concluimos que RKEF obtuvo muy buenos rendimientos. En particular, se observa una mejora significativa en *Ionosphere*, donde la dimensión se reduce bruscamente a una única componente. Esta situación favorece notablemente a RKEF, debido a que $d_2 = r - 1$ solo depende de la cantidad de clases. Por último, destacamos el hecho de que en los dos conjuntos de mayor dimensión (*Ionosphere* y *Satlog*), nuestro método logra significativamente los mejores resultados.

5.3 Comentarios de cierre de capítulo

En este capítulo presentamos una evaluación de la estrategia de reducción dimensional introducida en el Capítulo 4, usando datos simulados y reales con diferentes grados de dificultad. Pudimos ver que el método propuesto RKEF ofrece resultados que siempre están entre los mejores para cada experimento. Destacamos que las mayores ventajas se logran con datos reales de mayor dimensión, como los datos de microbioma analizados a nivel de género (L6) y los datos de cáncer de páncreas. Asimismo, resaltamos que las tasas de error conseguidas con RKEF usando una dimensión estándar $d_2 = r-1$ son, en general, muy similares a las obtenidas estimando la dimensión óptima de la reducción, la cual resulta generalmente mayor que $r - 1$. Por esta razón, parece poco significativo estimar dicha dimensión óptima, al menos desde un punto de vista práctico. Otro aspecto a remarcar es la tendencia a evitar el sobreajuste cuando se utilizan heurísticas típicas para escoger los parámetros de los núcleos, lo cual también resulta relevante en las aplicaciones.

CAPÍTULO **6**

Reducción suficiente de dimensiones con información adicional

En algunas situaciones, además de obtener muestras de la variable predictora $\mathbf{X} \in \mathbb{R}^p$ para construir un método de predicción para la variable respuesta $Y \in \mathbb{R}$, es posible medir en primera instancia una variable adicional $\mathbf{W}$ que posiblemente contenga buena información acerca de Y. No obstante, generalmente $\mathbf{W}$ es más costosa de obtener que $\mathbf{X}$, por lo cual quisiéramos evitar medirla en el futuro. Por lo tanto, estudiar directamente el problema $Y|(\mathbf{X}, \mathbf{W})$ no es recomendable. Por otro lado, si la información que tiene $\mathbf{W}$ sobre Y es valiosa, estudiar únicamente el problema $Y|\mathbf{X}$ podría ser poco provechoso. Luego, es de especial interés buscar una estrategia para obtener una reducción $\mathbf{R}(\mathbf{X})$ que no solo retenga la información que tiene $\mathbf{X}$ sobre Y, sino que además capture la información adicional provista por $\mathbf{W}$. Para conseguir tal reducción, queremos involucrar a $\mathbf{W}$ únicamente durante el proceso de entrenamiento.

Para motivar la idea de este capítulo, veamos el siguiente ejemplo que muestra la posible ventaja de utilizar también la información adicional para estudiar $Y|\mathbf{X}$.

Ejemplo 6.1. [Hung et al., 2015, Ejemplo 1.1] Sea $W \in \mathbb{R}$ y asumamos las distribuciones condicionales $W|\mathbf{X} \sim \mathcal{N}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}, 1 - b^2)$ y $Y|(\mathbf{X}, W) \sim \mathcal{N}(\boldsymbol{\gamma}^{\mathrm{T}}\mathbf{X} + aW, \sigma^2)$, con $\boldsymbol{\beta} \in \mathbb{R}^p$, $b = \|\boldsymbol{\beta}\|_2 < 1$, $\boldsymbol{\gamma} \in \mathbb{R}^p$ y $a \in \mathbb{R}^+$. Observar que a regula la influencia de W en Y, mientras

que b controla la correlación entre $\mathbf{X}$ y W. Entonces, fijando $\mathbf{X}$ y aplicando la Proposición A.1, resulta

$$Y|\mathbf{X} \sim \mathcal{N}\left((a\boldsymbol{\beta}+\boldsymbol{\gamma})^{\mathrm{T}}\mathbf{X}, \sigma^2 + a^2(1-b^2)\right). \tag{6.1}$$

La ecuación (6.1) nos indica que predecir Y sin la información adicional W incrementa $a^2(1-b^2)$ la varianza de Y. En consecuencia, a medida que el valor de a aumenta y el valor de b disminuye, se vuelve más ineficiente usar únicamente $\mathbf{X}$ para extraer información acerca de Y. □

El uso de métodos clásicos de reducción lineal para construir una estrategia de incorporación de la información adicional ya ha sido explorado en [Hung et al., 2015], donde los autores presentan un *método en dos pasos* que combina técnicas de reducción existentes. Sin embargo, dicha metodología tiene las limitaciones propias de considerar reducciones lineales en un ambiente de baja dimensión, por lo cual es importante estudiar cómo extender esta idea a escenarios de mayor complejidad.

Iniciaremos este capítulo resumiendo en la Sección 6.1 el concepto de SDR parcial, una propuesta de [Chiaromonte et al., 2002] que involucra el uso de una variable cualitativa $W \in \mathbb{R}$ durante la aplicación de los métodos clásicos de SDR. Este nuevo escenario involucra diferentes subespacios de reducción, cuyas relaciones estableceremos en la Sección 6.2. Luego, en la Sección 6.3 repasaremos el método de reducción en dos pasos de [Hung et al., 2015] y, posteriormente, en la Sección 6.4 propondremos una metodología más generalizada que permita explotar el potencial de los métodos de reducción no lineal basados en núcleos. Finalizaremos con ejemplos con datos reales en la Sección 6.5.

6.1 Reducción suficiente parcial

En lo que sigue, consideraremos $\mathbf{X} \in \mathcal{X} \subset \mathbb{R}^p$, $Y \in \mathcal{Y} \subset \mathbb{R}$ y $W \in \mathcal{W} = \{1, \ldots, s\}$. Denotaremos $(\mathbf{X}_w, Y_w)$ para indicar las variables sujetas a un w fijo, es decir, para referirnos a $(\mathbf{X}, Y)|(W = w)$. Además, escribiremos $\boldsymbol{\mu}_w = \mathbb{E}[\mathbf{X}_w]$ y $\boldsymbol{\Sigma}_w = \mathrm{var}(\mathbf{X}_w)$. Por último, denotaremos $\bigoplus_{i\in I} \mathcal{S}_i$ la suma directa de los subespacios $\mathcal{S}_i$.

La primera extensión de la teoría de SDR a situaciones que involucran dos grupos de variables predictoras fue presentada en [Chiaromonte et al., 2002]. El objetivo de los autores fue extender la teoría clásica de SDR y el método de estimación SIR [Li, 1991] a problemas de regresión que involucran variables tanto cuantitativas como categóricas. Esto es, considerar la regresión de Y en $(\mathbf{X}, W)$, donde $\mathbf{X}$ es la variable predictora cuantitativa y W corresponde a alguno de los siguientes casos: una variable cualitativa, una combinación de variables cualitativas o la categorización de una variable continua que contiene información cualitativa.

Es importante remarcar que, en el contexto indicado, no se pensaba específicamente que W fuese una variable difícil de medir en el futuro, sino que se pretendía simplificar el problema de la regresión $Y|(\mathbf{X}, W)$ a subproblemas donde los métodos clásicos fuesen aplicables (en ese momento los métodos existentes requerían $\mathbf{X}$ continua), pagando el costo de obtener una reducción que dependa únicamente de $\mathbf{X}$. Es decir, consideran el problema de regresión $Y|(\mathbf{X}, W)$ y plantean como objetivo hallar una transformación $\mathbf{R} : \mathbb{R}^p \to \mathbb{R}^d$ que verifique

$$Y \perp\!\!\!\perp \mathbf{X}|(\mathbf{R}(\mathbf{X}), W). \tag{6.2}$$

En particular, para el caso de SDR lineal, (6.2) se reescribe como

$$Y \perp\!\!\!\perp \mathbf{X}|(\mathbf{P}_{\mathcal{S}}\mathbf{X}, W). \tag{6.3}$$

Definición 6.2. [Chiaromonte et al., 2002] Sea $\mathcal{S}^{(W)}_{Y|\mathbf{X}}$ la intersección de todos los subespacios $\mathcal{S}$ que verifican (6.3). Si $\mathcal{S}^{(W)}_{Y|\mathbf{X}}$ existe y también verifica (6.3), se denomina *subespacio central parcial* (PCS) relativo a $\mathbf{X}$ para la regresión de Y en $(\mathbf{X}, W)$. Por su parte, $d^{(W)}_{Y|\mathbf{X}} := \dim(\mathcal{S}^{(W)}_{Y|\mathbf{X}})$ es la *dimensión estructural parcial*.

Ahora bien, fijando $w \in \mathcal{W}$, se puede considerar el problema $Y|(\mathbf{X}, W = w)$, correspondiente a la regresión de Y_w en $\mathbf{X}_w$. Si denotamos $\mathcal{S}_{Y_w|\mathbf{X}_w}$ el subespacio central y hacemos variar w, estamos considerando s sub-problemas de SDR y, en consecuencia, s subespacios centrales. La relación entre dichos subespacios centrales y el subespacio central parcial $\mathcal{S}^{(W)}_{Y|\mathbf{X}}$ de la Definición 6.2 queda establecida por el siguiente resultado.

Proposición 6.3. [Chiaromonte et al., 2002, Proposición 3.3] *Se verifica*

$$\mathcal{S}_{Y|\mathbf{X}}^{(W)} = \bigoplus_{w=1}^{s} \mathcal{S}_{Y_w|\mathbf{X}_w}. \tag{6.4}$$

A continuación, supongamos que la matriz de covarianza de $\mathbf{X}_w$ es común para todos los subproblemas. Es decir,

$$\boldsymbol{\Sigma}_w = \boldsymbol{\Sigma}_{\text{pool}}, \qquad \forall w \in \mathcal{W}. \tag{6.5}$$

Entonces, podemos considerar $\mathbf{Z}_w := \boldsymbol{\Sigma}_{\text{pool}}^{-1/2}(\mathbf{X}_w - \boldsymbol{\mu}_w)$ tal que, teniendo en cuenta la Proposición 3.10, la expresión (6.4) se reescribe como

$$\mathcal{S}_{Y|\mathbf{X}}^{(W)} = \bigoplus_{w=1}^{s} \boldsymbol{\Sigma}_w^{-1/2} \mathcal{S}_{Y_w|\mathbf{Z}_w} = \boldsymbol{\Sigma}_{\text{pool}}^{-1/2} \bigoplus_{w=1}^{s} \mathcal{S}_{Y_w|\mathbf{Z}_w}. \tag{6.6}$$

O lo que es lo mismo,

$$\mathcal{S}_{Y|\mathbf{X}}^{(W)} = \boldsymbol{\Sigma}_{\text{pool}}^{-1/2} \mathcal{S}_{Y|\mathbf{Z}}^{(W)} \tag{6.7}$$

Debido a que el objetivo es estimar el PCS $\mathcal{S}_{Y|\mathbf{X}}^{(W)}$, la relación (6.6) permite descomponer dicho subespacio y abordar cada uno de los subproblemas mediante algún método clásico de reducción.

6.1.1 Partial Sliced Inverse Regression (PSIR) [Chiaromonte et al., 2002]

En virtud de la relación (6.6), una opción es aplicar SIR a cada uno de los subproblemas. Para ello, siguiendo la notación de la Subsección 3.2.1, se define

$$\mathbf{K}_{\text{SIR}}^{w} := \text{var}\left(\mathbb{E}\left[\mathbf{Z}_w | Y_w\right]\right) \tag{6.8}$$

y se promedia sobre W para obtener

$$\mathbf{K}_{\text{SIR}}^{(W)} := \sum_{w=1}^{s} \mathbb{P}\left\{W = w\right\} \mathbf{K}_{\text{SIR}}^{w}. \tag{6.9}$$

De manera análoga al Corolario 3.12, se presenta el siguiente resultado.

Proposición 6.4. [Chiaromonte et al., 2002, Proposición 4.1] *Bajo (6.5) y la condición de linealidad* $\mathbb{E}\left[\mathbf{Z}_w|\mathbf{P}_{\mathcal{S}_{Y_w|\mathbf{Z}_w}}\mathbf{Z}_w\right] = \mathbf{P}_{\mathcal{S}_{Y_w|\mathbf{Z}_w}}\mathbf{Z}_w$ *para todo* $w \in \mathcal{W}$*, se verifica*

$$\operatorname{span}\mathbf{K}_{\mathrm{SIR}}^{(W)} \subset \bigoplus_{w=1}^{s} \mathcal{S}_{Y_w|\mathbf{Z}_w}.$$

En virtud de la Proposición 6.4 y de la relación (6.7), en [Chiaromonte et al., 2002] proponen un método de estimación de $\mathcal{S}_{Y|\mathbf{X}}^{(W)}$, el cual presentamos a continuación.

Método de estimación **PSIR**

---

1. Para cada $w \in \mathcal{W}$, aplicar SIR para obtener la estimación $\hat{\mathbf{K}}_{\mathrm{SIR}}^{w}$ de (6.8).

2. Estimar $\mathbf{K}_{\mathrm{SIR}}^{(W)}$ en (6.9) mediante

$$\hat{\mathbf{K}}_{\mathrm{SIR}}^{(W)} = \sum_{w=1}^{s} \frac{n_w}{n} \hat{\mathbf{K}}_{\mathrm{SIR}}^{w},$$

   donde n es el tamaño muestral y n_w es el tamaño de la submuestra correspondiente a $W = w$.

3. Estimar $\mathcal{S}_{Y|\mathbf{Z}}^{(W)}$ a partir de los autovectores correspondientes a los d autovalores más grandes de $\hat{\mathbf{K}}_{\mathrm{SIR}}^{(W)}$.

4. Finalmente, estimar $\mathcal{S}_{Y|\mathbf{X}}^{(W)}$ a partir de la relación (6.7), usando la versión muestral de $\mathbf{\Sigma}_{\mathrm{pool}}$ dada por

$$\hat{\mathbf{\Sigma}}_{\mathrm{pool}} = \sum_{w=1}^{s} \frac{n_w}{n} \hat{\mathbf{\Sigma}}_w.$$

6.2 Relaciones entre subespacios de reducción

Supongamos $W \in \mathbb{R}$ y sea el problema de reducción lineal

$$Y \perp\!\!\!\perp \mathbf{X}|\mathbf{\Gamma}^{\mathrm{T}}\mathbf{X} \qquad \text{con información adicional } W,$$

donde $\mathbf{\Gamma} \in \mathbb{R}^{p\times d}$ verifica $\operatorname{span}\mathbf{\Gamma} = \mathcal{S}_{Y|\mathbf{X}}$. Estableceremos un par de relaciones entre:

- El subespacio central $\mathcal{S}_{Y|\mathbf{X}}$ para la regresión de Y en $\mathbf{X}$.
- El subespacio central $\mathcal{S}_{W|\mathbf{X}}$ para la regresión de W en $\mathbf{X}$.
- El subespacio central $\mathcal{S}_{(Y,W)|\mathbf{X}}$ para la regresión de (Y, W) en $\mathbf{X}$.
- El subespacio central parcial $\mathcal{S}^{(W)}_{Y|\mathbf{X}}$ relativo a $\mathbf{X}$ para la regresión de Y en $(\mathbf{X}, W)$.

Por Proposición 3.10, basta trabajar con $\mathbf{Z} := \boldsymbol{\Sigma}^{-1/2}(\mathbf{X} - \boldsymbol{\mu})$. La idea de incorporar W al estudio se basa en el siguiente resultado, cuya prueba se presenta en el Anexo D.1.

Proposición 6.5. *Se verifica*

$$\mathcal{S}_{Y|\mathbf{Z}} \subset \mathcal{S}_{(Y,W)|\mathbf{Z}}. \tag{6.10}$$

La relación (6.10) indica que $\mathcal{S}_{(Y,W)|\mathbf{Z}}$ encapsula al subespacio de interés $\mathcal{S}_{Y|\mathbf{Z}}$, por lo cual la búsqueda de las proyecciones de $\mathbf{Z}$ a partir de métodos de reducción lineal puede restringirse para hacer más efectiva la estimación. Veremos esto en detalle en la siguiente sección. Ahora, dada la importancia de $\mathcal{S}_{(Y,W)|\mathbf{Z}}$, introduciremos la siguiente definición.

Definición 6.6. [Hung et al., 2015, Definición 3.1] El subespacio $\mathcal{S}_{(Y,W)|\mathbf{Z}}$ se denomina *subespacio W-envolvente del subespacio central* $\mathcal{S}_{Y|\mathbf{Z}}$. Denotamos

$$\mathcal{S}_{\text{env}} := \mathcal{S}_{(Y,W)|\mathbf{Z}}. \tag{6.11}$$

En lo que sigue, la dimensión $d_{\text{env}} := \dim(\mathcal{S}_{\text{env}})$ se supondrá conocida. La siguiente relación permite determinar $\mathcal{S}_{\text{env}}$ a partir de los subespacios $\mathcal{S}_{W|\mathbf{Z}}$ y $\mathcal{S}^{(W)}_{Y|\mathbf{Z}}$.

Proposición 6.7. [Hung et al., 2015, Proposición 3.1] *Se verifica*

$$\mathcal{S}_{\text{env}} = \mathcal{S}_{W|\mathbf{Z}} \oplus \mathcal{S}^{(W)}_{Y|\mathbf{Z}}. \tag{6.12}$$

A modo de ejemplo, analicemos las relaciones (6.10) y (6.12) en el Ejemplo 6.1, el cual presentamos como motivación al principio del capítulo.

Ejemplo 6.1 (continuación). De (6.1), se deduce que

$$\mathcal{S}_{Y|\mathbf{X}} = \operatorname{span}(a\boldsymbol{\beta} + \boldsymbol{\gamma}). \tag{6.13}$$

Del mismo modo, observando las distribuciones de $W|\mathbf{X}$ y $Y|(\mathbf{X}, W)$, está claro que $\mathcal{S}_{W|\mathbf{X}} = \operatorname{span}\boldsymbol{\beta}$ y $\mathcal{S}^{(W)}_{Y|\mathbf{X}} = \operatorname{span}\boldsymbol{\gamma}$. Por lo tanto, de (6.12) resulta

$$\mathcal{S}_{\text{env}} = \operatorname{span}\{\boldsymbol{\beta}, \boldsymbol{\gamma}\}. \tag{6.14}$$

Como $a\boldsymbol{\beta} + \boldsymbol{\gamma} \in \operatorname{span}\{\boldsymbol{\beta}, \boldsymbol{\gamma}\}$, de (6.13) y (6.14) se verifica la relación (6.10). □

6.3 Método lineal en dos pasos (HTS) [Hung et al., 2015]

Sea $\mathbf{K}_{Y|\mathbf{Z}}$ una matriz núcleo que satisface $\mathcal{S}_{Y|\mathbf{Z}} = \operatorname{span}\mathbf{K}_{Y|\mathbf{Z}}$, asociada a algún metodo clásico de SDR para la regresión de Y en $\mathbf{Z}$. Por ejemplo, para SIR [Li, 1991] y SAVE [Cook and Weisberg, 1991] se definen $\mathbf{K}_{\text{SIR}} := \operatorname{cov}(\mathbb{E}[\mathbf{Z}|Y])$ y $\mathbf{K}_{\text{SAVE}} := \mathbb{E}[\mathbf{I}_p - \operatorname{cov}(\mathbf{Z}|Y)]^2$, respectivamente. Si $\hat{\mathbf{K}}_{Y|\mathbf{Z}}$ es un estimador consistente de $\mathbf{K}_{Y|\mathbf{Z}}$, la estimación de una base para $\mathcal{S}_{Y|\mathbf{Z}}$ se realiza tradicionalmente a partir del problema de optimización

$$\max_{\substack{\{\boldsymbol{\beta}_1,\ldots,\boldsymbol{\beta}_d\}:\\ \|\boldsymbol{\beta}_i\|_2=1\\ \boldsymbol{\beta}_i^{\mathrm{T}}\boldsymbol{\beta}_j=0,\, i\neq j}} \sum_{k=1}^{d} \boldsymbol{\beta}_k^{\mathrm{T}}\, \hat{\mathbf{K}}_{Y|\mathbf{Z}}\, \boldsymbol{\beta}_k. \tag{6.15}$$

Para incorporar la información adicional W, en [Hung et al., 2015] proponen restringir la búsqueda de los $\boldsymbol{\beta}_k$ al subespacio W-envolvente $\mathcal{S}_{\text{env}}$. Es decir, reformulan (6.15) imponiendo $\{\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_d\} \subset \mathcal{S}_{\text{env}}$, lo cual deriva en un nuevo problema de optimización:

$$\max_{\substack{\{\boldsymbol{\beta}_1,\ldots,\boldsymbol{\beta}_d\}\subset\mathcal{S}_{\text{env}}:\\ \|\boldsymbol{\beta}_s\|_2=1\\ \boldsymbol{\beta}_s^{\mathrm{T}}\boldsymbol{\beta}_l=0, s\neq l}} \sum_{k=1}^{d} \boldsymbol{\beta}_k^{\mathrm{T}}\, \hat{\mathbf{K}}_{Y|\mathbf{Z}}\, \boldsymbol{\beta}_k. \tag{6.16}$$

Si $\mathbf{B}_{\text{env}}$ es una base de $\mathcal{S}_{\text{env}}$, la solución de (6.16) la conforman los autovectores correspondientes a los d autovalores más grandes de $\mathbf{P}_{\mathbf{B}_{\text{env}}}\hat{\mathbf{K}}_{Y|\mathbf{Z}}\mathbf{P}_{\mathbf{B}_{\text{env}}}$ [Naik and Tsai, 2005, Proposición 3], donde $\mathbf{P}_{\mathbf{B}_{\text{env}}}$ es la matriz de proyección a $\mathcal{S}_{\text{env}}$. Sin embargo, $\mathcal{S}_{\text{env}}$ es en general desconocido, por lo que es necesario estimar la base $\mathbf{B}_{\text{env}}$. Por esta razón, el método de [Hung et al., 2015], al cual denominaremos HTS, consiste en *dos pasos*:

1°) Estimación de la base $\mathbf{B}_{\text{env}}$ de $\mathcal{S}_{\text{env}}$.

2°) Estimación de la base $\mathbf{B}$ del subespacio de interés $\mathcal{S}_{Y|\mathbf{Z}}$ a partir de

$$\mathbf{P}_{\hat{\mathbf{B}}_{\text{env}}}\hat{\mathbf{K}}_{Y|\mathbf{Z}}\mathbf{P}_{\hat{\mathbf{B}}_{\text{env}}}. \tag{6.17}$$

El siguiente resultado indica bajo qué condiciones es posible asegurar la consistencia del método en dos pasos de Hung et al. [2015].

Teorema 6.8. [Hung et al., 2015, Teorema 3.2] *Sean* $\mathbf{K}_{Y|\mathbf{Z}}$ *y* $\mathbf{K}_{\text{env}}$ *matrices núcleos tales que* $\mathcal{S}_{Y|\mathbf{Z}} = \operatorname{span}\mathbf{K}_{Y|\mathbf{Z}}$ *y* $\mathcal{S}_{\text{env}} = \operatorname{span}\mathbf{K}_{\text{env}}$*, y sea* $\hat{\mathbf{B}}_{\text{env}}$ *la matriz formada por los autovectores correspondientes a los* ν *autovalores más grandes de* $\hat{\mathbf{K}}_{\text{env}}$*, con* $\nu \geq d_{\text{env}}$*. Si* $\hat{\mathbf{K}}_{Y|\mathbf{Z}}$ *y* $\hat{\mathbf{K}}_{\text{env}}$ *son estimadores consistentes de* $\mathbf{K}_{Y|\mathbf{Z}}$ *y* $\mathbf{K}_{env}$*, respectivamente, entonces (6.17) es un estimador consistente de* $\mathbf{K}_{Y|\mathbf{Z}}$*.*

Observar que la selección del parámetro ν depende de la estimación de la dimensión d_{env}, puesto que es necesario restringir la búsqueda a $\nu \geq d_{\text{env}}$ para que el método sea consistente, en virtud del Teorema 6.8. Más adelante, repasaremos qué criterios de selección proponen los autores de [Hung et al., 2015], tanto para las dimensiones d y d_{env} (Subsección 6.3.1), como para los parámetros ν y ω (Subsección 6.3.2).

Por otra parte, para la estimación de $\mathbf{K}_{\text{env}}$ existen dos opciones, dependiendo de qué expresión de $\mathcal{S}_{\text{env}}$ se usa. La primera opción es utilizar su definición en (6.11), la cual es directa pero implica tratar con un problema de regresión de respuesta multivariada (Y, W). La segunda opción es utilizar la relación (6.12), que permite descomponer $\mathcal{S}_{\text{env}}$ e incorporar la idea de importancia relativa entre $\mathcal{S}_{W|\mathbf{Z}}$ y $\mathcal{S}^{(W)}_{Y|\mathbf{Z}}$. Para ilustrar esto último, volvamos a utilizar la situación planteada en el Ejemplo 6.1.

Ejemplo 6.1 (continuación). Teniendo en cuenta que $\mathcal{S}_{Y|\mathbf{X}} = \operatorname{span} a\boldsymbol{\beta} + \boldsymbol{\gamma}$, $\mathcal{S}_{W|\mathbf{X}} = \operatorname{span}\boldsymbol{\beta}$ y $\mathcal{S}^{(W)}_{Y|\mathbf{X}} = \operatorname{span}\boldsymbol{\gamma}$, podemos observar que si a es pequeño, entonces $a\boldsymbol{\beta} + \boldsymbol{\gamma} \simeq \boldsymbol{\gamma}$. Esto significa que $\mathcal{S}^{(W)}_{Y|\mathbf{X}}$ influye más que $\mathcal{S}_{W|\mathbf{Z}}$ en la construcción de $\mathcal{S}_{\text{env}}$. □

Si bien tanto (6.11) como (6.12) conducen a definiciones de estimadores de $\mathbf{K}_{\text{env}}$, en lo que sigue expondremos solo la segunda opción, debido a que es más óptima para estimar $\mathcal{S}_{Y|\mathbf{Z}}$. El uso de (6.12) para proponer un estimador de $\mathbf{K}_{\text{env}}$ tiene como base el siguiente resultado, que es una consecuencia directa de la Proposición 6.7.

Corolario 6.9. *Sean* $\mathbf{K}_{Y|\mathbf{Z}}$ *y* $\mathbf{K}_{\text{env}}$ *matrices núcleos tales que* $\mathcal{S}_{Y|\mathbf{Z}} = \operatorname{span}\mathbf{K}_{Y|\mathbf{Z}}$ *y* $\mathcal{S}_{\text{env}} = \operatorname{span}\mathbf{K}_{\text{env}}$. *Para todo* $\omega \in (0,1)$, *se verifica*

$$\operatorname{span}\left(\omega\mathbf{K}_{W|\mathbf{Z}} + (1-\omega)\mathbf{K}_{Y|\mathbf{Z}}^{(W)}\right) = \mathcal{S}_{\text{env}}.$$

Método de estimación **HTS**

---

1. Construir $\hat{\mathbf{B}}_{\text{env}}$ a partir de los autovectores correspondientes a los ν autovalores más grandes de

$$\hat{\mathbf{K}}_{\text{env}}(\omega) := \omega\hat{\mathbf{K}}_{W|\mathbf{Z}} + (1-\omega)\hat{\mathbf{K}}_{Y|\mathbf{Z}}^{(W)}, \qquad \omega \in (0,1). \tag{6.18}$$

2. Estimar $\mathcal{S}_{Y|\mathbf{Z}}$ a partir de la matriz $\hat{\mathbf{B}}(\nu,\omega)$ formada por los autovectores correspondientes a los d autovalores más grandes de (6.17).

Observar que el método en dos pasos HTS requiere la implementación tanto de métodos de reducción clásico, para $\hat{\mathbf{K}}_{W|\mathbf{Z}}$ y $\hat{\mathbf{K}}_{Y|\mathbf{Z}}$, como de algún método de reducción parcial para $\hat{\mathbf{K}}_{Y|\mathbf{Z}}^{(W)}$. Las herramientas sugeridas y utilizadas en [Hung et al., 2015] son SIR y SAVE, ambos métodos de reducción lineal, y el método de reducción parcial PSIR expuesto en la Subsección 6.1.1.

Además, es importante remarcar que el valor de $\omega \in (0,1)$ no afecta la consistencia de $\hat{\mathbf{K}}_{\text{env}}(\omega)$ como estimador de $\mathbf{K}_{\text{env}}$, en virtud del Corolario 6.9. Por lo tanto, el Teorema 6.8 es válido en la formulación de HTS, de manera que está asegurada la consistencia de la estimación de $\mathcal{S}_{Y|\mathbf{Z}}$.

6.3.1 Determinación de d y d_{env}

Sea $\omega \in (0,1)$. En [Hung et al., 2015] eligen determinar d y d_{env} mediante un criterio de tipo BIC propuesto en [Zhu et al., 2010], el cual deriva en las expresiones

$$\hat{d}_{\text{env}}(\omega) := \underset{k=1,\ldots,p}{\arg\max} \left\{ \frac{n \sum_{j=1}^{k} \left[\ln(\hat{\lambda}_j + 1) - \hat{\lambda}_j \right]}{2 \sum_{j=1}^{p} \left[\ln(\hat{\lambda}_j + 1) - \hat{\lambda}_j \right]} - 2C_n \frac{k(k-1)}{2p} \right\}$$

y

$$\hat{d}(\omega) := \underset{k=1,\ldots,\hat{d}_{\text{env}}}{\arg\max} \left\{ \frac{n \sum_{j=1}^{k} \left[\ln(\hat{\lambda}_j^* + 1) - \hat{\lambda}_j^* \right]}{2 \sum_{j=1}^{p} \left[\ln(\hat{\lambda}_j^* + 1) - \hat{\lambda}_j^* \right]} - 2C_n \frac{k(k-1)}{2p} \right\},$$

donde $\left\{ \hat{\lambda}_j : j = 1, \ldots, p \right\}$ y $\left\{ \hat{\lambda}_j^* : j = 1, \ldots, p \right\}$ son los autovalores de $\hat{\mathbf{K}}_{\text{env}}$ y (6.17), respectivamente, y C_n es un parámetro de penalización. Si C_n cumple con la condición de que $C_n/n \to 0$ y $C_n \to \infty$ cuando $n \to \infty$, entonces $\hat{d}_{\text{env}}(\omega)$ y $\hat{d}(\omega)$ son estimadores consistentes [ver Zhu et al., 2010, Teorema 4]. Finalmente, para un conjunto finito $\Omega \subset (0,1)$, se eligen

$$\hat{d}_{\text{env}} = \text{median}\left\{ \hat{d}_{\text{env}}(\omega) : \omega \in \Omega \right\} \qquad \text{y} \qquad \hat{d} = \text{median}\left\{ \hat{d}(\omega) : \omega \in \Omega \right\}.$$

6.3.2 Selección de parámetros

Recordemos que el Teorema 6.8 asegura la consistencia de HTS siempre que $\nu \geq d_{\text{env}}$. Además, por Proposición 6.9, queda claro que $\omega \in (0,1)$ no afecta la consistencia de dicho método. Por esa razón, suponiendo que $\hat{d}_{\text{env}} \geq d_{\text{env}}$, la principal influencia de (ν, ω) recae en la variación del estimador y, en consecuencia, en [Hung et al., 2015] proponen seleccionar ambos parámetros mediante el criterio de mínima variabilidad de [Ye and Weiss, 2003]. Este criterio implica minimizar respecto de (ν, ω) la variabilidad de $\hat{\mathbf{B}}$ definida por

$$v(\nu, \omega) := \frac{1}{m} \sum_{b=1}^{m} q\left(\hat{\mathbf{B}}^{(b)}(\nu, \omega), \hat{\mathbf{B}}(\nu, \omega) \right),$$

donde $q(\mathbf{B}_1, \mathbf{B}_2) := \|\mathbf{P}_{\mathbf{B}_1} - \mathbf{P}_{\mathbf{B}_2}\|_{\text{F}}^2$, con $\|\cdot\|_{\text{F}}$ la norma de Frobenius, y $\left\{ \hat{\mathbf{B}}^{(b)} : b = 1, \ldots, m \right\}$ es una familia de estimadores a dos pasos obtenida mediante bootstrap. Observar que este criterio depende de los métodos de reducción involucrados. Por supuesto, en caso de

evaluar el rendimiento de algún algoritmo de regresión o clasificación, según corresponda, se puede optar por seleccionar los parámetros mediante un proceso de validación cruzada.

6.4 Método en dos pasos generalizado

Consideraremos ahora $\mathbf{W} \in \mathbb{R}^m$ y el problema de reducción general

$$Y \perp\!\!\!\perp \mathbf{X} | \mathbf{R}(\mathbf{X}) \qquad \text{con información adicional } \mathbf{W}.$$

Dadas las limitaciones de aplicabilidad del método en dos pasos de [Hung et al., 2015], buscamos una alternativa para poder generalizar la idea, de tal manera de poder elegir más libremente los métodos de reducción en el proceso. Para ello, necesitamos la siguiente versión general de la Proposición 6.5. La demostración se presenta en el Anexo D.2.

Proposición 6.10. *Si* $\mathbf{R}(\mathbf{X})$ *es una SDR de* $\mathbf{X}$ *para la regresión de* $(Y, \mathbf{W})$ *en* $\mathbf{X}$*, entonces también lo es para la regresión de* Y *en* $\mathbf{X}$*.*

El resultado anterior nos brinda una herramienta para incorporar la información adicional $\mathbf{W}$. De esta manera, estamos en condiciones de formalizar nuestro *método en dos pasos generalizado*, el cual conduce a una SDR de $\mathbf{X}$ para la regresión de Y en $\mathbf{X}$.

Método en dos pasos generalizado

(I) Obtener una SDR $\mathbf{R}^{(1)}(\mathbf{X})$ para la regresión de $(Y, \mathbf{W})$ en $\mathbf{X}$.

(II) Obtener una SDR $\mathbf{R}(\mathbf{X}) = \mathbf{R}^{(2)}(\mathbf{R}^{(1)}(\mathbf{X}))$ para la regresión de Y en $\mathbf{R}^{(1)}(\mathbf{X})$.

El siguiente teorema nos asegura la suficiencia de la reducción propuesta por el método en dos pasos generalizado, y además muestra que la misma es independiente del modelo adoptado para la regresión inversa $\mathbf{X}|Y$. La demostración se encuentra en el Anexo D.3.

Teorema 6.11 . *Si* $\mathbf{R}^{(1)}(\mathbf{X})$ *es una SDR de* $\mathbf{X}$ *para la regresión de* $(Y, \mathbf{W})$ *en* $\mathbf{X}$ *y* $\mathbf{R}^{(2)}(\mathbf{R}^{(1)}(\mathbf{X}))$ *es una SDR de* $\mathbf{R}^{(1)}(\mathbf{X})$ *para la regresión de* Y *en* $\mathbf{R}^{(1)}(\mathbf{X})$*, entonces*

$$\mathbf{R}(\mathbf{X}) = (\mathbf{R}^{(2)} \circ \mathbf{R}^{(1)})(\mathbf{X}) \tag{6.19}$$

es una SDR de $\mathbf{X}$ *para la regresión de* Y *en* $\mathbf{X}$*.*

El método en dos pasos generalizado nos deja a libre elección los métodos a utilizar para las reducciones $\mathbf{R}^{(1)}$ y $\mathbf{R}^{(2)}$, pudiéndose incluso combinar distintos métodos. Utilizaremos la siguiente notación:

- Si $\mathbf{R}^{(1)}$ y $\mathbf{R}^{(2)}$ se obtienen con los métodos MA y MB, respectivamente, denominaremos a la reducción $\mathbf{R}(\mathbf{X})$ de (6.19) como MB-AI(MA). Por ejemplo, RKEF-AI(COIR) implica implementar COIR en el paso (I) y luego RKEF en el (II).
- Si MA y MB son el mismo método, denominaremos a $\mathbf{R}(\mathbf{X})$ directamente MA-AI. Por ejemplo, RKEF-AI implica implementar RKEF en ambos pasos.

Por otra parte, el método en dos pasos generalizado no realiza suposiciones sobre las variables involucradas, permitiéndonos implementarlo aún si $\mathbf{W}$ es multivariada. Por supuesto, la naturaleza de las variables nos orientará sobre qué métodos de reducción usar, especialmente en el paso interno de obtener $\mathbf{R}^{(1)}(\mathbf{X})$, que es donde se incorpora $\mathbf{W}$.

Así, para el caso multivariado $\mathbf{W} \in \mathbb{R}^m$, nuestra recomendación es utilizar métodos basados en núcleos; en particular, la combinación RKEF-AI(COIR). La elección de COIR se debe a que es menos sensible al tipo de variable $(Y, \mathbf{W})$, dado que en la práctica se reconocen ciertos núcleos apropiados para los diferentes casos. Luego, una vez incorporada la información adicional $\mathbf{W}$, consideramos que nuestro método de reducción RKEF tiene un buen desempeño, como vimos en la parte experimental durante el Capítulo 5.

Por último, si el problema solo involucra Y y W discretas, y además las cantidades r y s no son lo suficientemente grandes, entonces en ambos pasos del método resulta práctico aplicar RKEF. Esto deriva en RKEF-AI, un caso particular del método en dos pasos generalizado, que implica trabajar durante todo el proceso en un contexto de KEF. Por ese motivo, lo analizaremos a continuación con más detalle.

6.4.1 Caso especial: método en dos pasos vía RKEF

Supongamos $Y \in \mathcal{Y} = \{1, \ldots, r\}$ y $W \in \mathcal{W} = \{1, \ldots, s\}$. Si modelamos tanto $\mathbf{X}|Y$ como $\mathbf{X}|(Y, W)$ mediante una KEF, podemos utilizar nuestra propuesta para construir sucesivamente $\mathbf{R}^{(1)}(\mathbf{X})$ y $\mathbf{R}^{(2)}(\mathbf{R}^{(1)}(\mathbf{X}))$ en el Teorema 6.11.

En el caso del problema de clasificación $(Y, W)|\mathbf{X}$, lo simplificaremos definiendo una variable discreta M que será la indexación de (Y, W), y que estará formada por todas las combinaciones posibles entre Y y W; es decir,

$$M \in \mathcal{M} \subset \{(i-1)s + j : i \in \mathcal{Y}, j \in \mathcal{W}\}. \tag{6.20}$$

Resulta $\mathcal{M} = \{1, \ldots, m\}$, con $m \leq rs$. Una consecuencia de los Teoremas 4.1 y 6.11 es el siguiente resultado, cuya demostración se encuentra en el Anexo D.4. Por supuesto, una versión análoga se consigue utilizando el Teorema 4.8 en lugar del Teorema 4.1.

Teorema 6.12. *Para (Y, W), sea $M \in \mathcal{M} = \{1, \ldots, m\}$ la indexación dada por (6.20). Sean $\mathbf{R}^{(1)} : \mathbb{R}^p \to \mathbb{R}^{m-1}$ y $\mathbf{R}^{(2)} : \mathbb{R}^{m-1} \to \mathbb{R}^{r-1}$ las SDR (4.3) de $\mathbf{X}$ para $M|\mathbf{X}$ y $Y|\mathbf{R}^{(1)}(\mathbf{X})$, respectivamente, y supongamos que las condiciones del Teorema 4.1 se verifican en cada paso. La transformación $\mathbf{R} : \mathbb{R}^p \to \mathbb{R}^{r-1}$ dada por*

$$\mathbf{R}(\mathbf{X}) = (\mathbf{R}^{(2)} \circ \mathbf{R}^{(1)})(\mathbf{X})$$

es una SDR de $\mathbf{X}$ para el problema de clasificación $Y|\mathbf{X}$.

Sea $\mathcal{D}_{(\mathbf{x},w,y)}$ un conjunto de datos de $(\mathbf{X}, W, Y)$, a continuación definimos la reducción RKEF-AI, en concordancia con la Definición 4.14 de la reducción RKEF.

Definición 6.13. Sean $\hat{\mathbf{R}}^{(1)}_{\text{SVM}} : \mathbb{R}^p \to \mathbb{R}^{m(m-1)/2}$ y $\hat{\mathbf{R}}^{(2)}_{\text{SVM}} : \mathbb{R}^{m(m-1)/2} \to \mathbb{R}^{r(r-1)/2}$ las reducciones RKEF definidas por (4.14) para $M|\mathbf{X}$ y $Y|\hat{\mathbf{R}}^{(1)}_{\text{SVM}}(\mathbf{X})$, respectivamente. Definimos la *reducción en KEF restringida vía SVM con información adicional* (RKEF-AI) como

$$\hat{\mathbf{R}}_{\text{SVM-AI}}(\mathbf{X}) = (\hat{\mathbf{R}}^{(2)}_{\text{SVM}} \circ \hat{\mathbf{R}}^{(1)}_{\text{SVM}})(\mathbf{X}).$$

Finalmente, estamos en condiciones de utilizar nuestro método RKEF, formalizado en la Sección 4.5, en ambas etapas del método en dos pasos generalizado. Esto deriva en lo que denominaremos *método en dos pasos vía RKEF* (o simplemente, RKEF-AI). En concordancia con RKEF, es posible que se obtengan reducciones menores a $m(m-1)/2$ y $r(r-1)/2$ en el proceso. De hecho, de acuerdo con lo expuesto anteriormente y con la experimentación del Capítulo 5, estableceremos en la próxima subsección una estrategia adecuada para la selección de d y d_{env}.

Por otra parte, es evidente que el método RKEF-AI tiene ciertas limitaciones. Si r o s son lo suficientemente grandes, pueden surgir inconvenientes en el primer paso, incluso cuando se limite el conjunto de búsqueda de d_{env}. Ese puede ser típicamente el escenario si Y o W son continuas, en cuyo caso r y/o s están determinados por el número de *slices* que se consideren en la discretización. En todos los casos donde RKEF-AI no sea apropiado, recomendamos aplicar RKEF-AI(COIR), como discutimos previamente.

6.4.2 Determinación de d y d_{env}

Del Teorema 6.12 se deduce que las dimensiones estructurales de los subespacios de reducción $\mathcal{S}_{\text{env}}$ y $\mathcal{S}_{Y|\mathbf{X}}$ verifican $d_{\text{env}} \leq \text{mín}\{m-1, p\}$ y $d \leq d_{\text{env}}$, respectivamente. Por supuesto, podemos añadir esa información a los métodos de reducción utilizados en el método en dos pasos generalizado, limitando los conjuntos de selección de ambas dimensiones. En consecuencia, y teniendo en cuenta que para evaluar nuestros métodos de reducción utilizaremos algoritmos de regresión o clasificación, según corresponda, optaremos por seleccionar $\hat{d}$ y $\hat{d}_{\text{env}}$ mediante validación cruzada, considerando los siguientes conjuntos de búsqueda:

$$\hat{d}_{\text{env}} \in \{1, \ldots, \text{mín}\{m-1, p\}\} \qquad \text{y} \qquad \hat{d} \in \{1, \ldots, \hat{d}_{\text{env}}\}. \tag{6.21}$$

Es decir, se procederá de manera similar a lo expuesto en la Subsección 4.3, utilizando como criterio el error de validación cruzada de algún algoritmo de clasificación para las diferentes combinaciones de $\hat{d}_{\text{env}}$ y $\hat{d}$ en (6.21).

	Variables		
Datos	Predictoras	Adicional	Respuesta
WPBC	$\mathbf{X} \in \mathbb{R}^{30}$	$W \in \{1, 2, 3\}$	$Y \in \{1, 2\}$
PanCancer	$\mathbf{X} \in \mathbb{R}^{1000}$	$W \in \mathbb{R}^{50}$	$Y \in \{1, \ldots, 10\}$

TABLA 6.1. Descripción de datos de Sección 6.5.

6.5 Ejemplos con datos reales

En esta sección evaluaremos el método en dos pasos generalizado en diferentes conjuntos de datos, resumidos en la Tabla 6.1. En primer lugar, en la Subsección 6.5.1, se analizará un conjunto de datos de baja dimensión donde la información adicional en W es discreta, escenario en el cual son aplicables tanto el método en dos pasos HTS de [Hung et al., 2015] como nuestra propuesta RKEF-AI. Luego, en la Subsección 6.5.2, se considerará un escenario no solamente de mayor dimensión sino con la información adicional $\mathbf{W}$ formada por un grupo de variables continuas, lo cual es tratado en forma particular por la combinación RKEF-AI(COIR). Por supuesto, pretendemos mostrar el potencial de la generalización del método en dos pasos en situaciones complejas.

En cada uno de los análisis es de especial interés incluir los resultados de estudiar de forma directa el problema $Y|\mathbf{X}$, ignorando la información adicional. En estos casos, usaremos nuestro método de reducción RKEF.

Como antes, utilizaremos como criterio de evaluación el error de clasificación de diferentes algoritmos. Por su parte, la determinación de $\hat{d}$ y $\hat{d}_{\text{env}}$ se realizará en concordancia con lo expuesto para cada método de reducción, mientras que la selección de parámetros de los clasificadores que lo requieran se realizará mediante alguno de los criterios definidos en la Sección 4.4.

6.5.1 Datos de cáncer de mamas

Los datos *breast cancer Wisconsin prognostic* (WPBC), disponibles en el repositorio UCI, consisten en 253 casos de cáncer de mamas. Además, corresponden a un problema

supervisado de clasificación binaria $Y|\mathbf{X}$, cuyo objetivo es predecir la recurrencia o no del cáncer a partir de 30 variables predictoras, calculadas mediante la imagen digitalizada de una muestra de células obtenidas por aspiración con aguja fina. Por lo tanto, resulta $\mathbf{X} \in \mathbb{R}^{30}$ y $Y \in \{0, 1\}$. La variable adicional $W \in \{0, 1, 2\}$ es el estado de la enfermedad, determinado por el tamaño del tumor y el estado de los ganglios linfáticos.

Dado que $r = 2$ y $s = 3$, resulta $m = 6$ y, como $m - 1 = 5 < 30 = p$, es apropiado aplicar RKEF-AI. Respecto a $\hat{d}$ y $\hat{d}_{\text{env}}$ en (6.21), la situación es la siguiente:

- Para el paso (I), $\hat{d}_{\text{env}} \in \{1, \ldots, 5\}$.
- Para el paso (II), al ser un problema de clasificación binaria, $\hat{d} = 1$.

Además, también es un buen escenario para el método HTS, por lo cual el principal propósito es comparar ambos métodos de reducción con información adicional.

Procedimiento *6.5.1. Datos WPBC*

1° Particionar los datos para efectuar validación cruzada 10-*fold*.

2° Obtener vía RKEF-AI la reducción $\hat{R}_{\text{SVM-AI}}$ de dimensión $\hat{d} = 1$, con $\hat{d}_{\text{env}} \in \{1, \ldots, 5\}$. Utilizar *Criterio 1* para seleccionar el ancho de banda σ del núcleo Gaussiano.

3° Para comparar con RKEF-AI, obtener la reducción $\hat{R}_{\text{HTS}}$ del método HTS, utilizando SIR y PSIR en el paso (I) y SAVE para $\hat{\mathbf{K}}_{Y|\mathbf{Z}}$ en el paso (II). Utilizar *Criterio 2* para $\omega \in \{0.1, 0.2, \ldots, 0.9\}$ en (6.18).

4° Obtener vía RKEF la reducción $\hat{R}_{\text{SVM}}$ sin información adicional, de dimensión $\hat{d} = 1$. Utilizar *Criterio 1* para seleccionar el ancho de banda σ del núcleo Gaussiano.

5° Calcular el error de clasificación 10-*fold* CV de los clasificadores LDA, QDA, LSVM y MLP. Utilizar *Criterio 2* para el parámetro de costo en LSVM y el número de neuronas en la capa oculta de MLP.

	RKEF	RKEF-AI		HTS	
Clasificador	Error	$\hat{d}_{\text{env}}$	Error	$\hat{d}_{\text{env}}$	Error
LDA	0.225 ± 0.060	5	$\mathbf{0.186 \pm 0.050}$	5	0.191 ± 0.023
QDA	0.260 ± 0.084	3	0.198 ± 0.026	3	$\mathbf{0.194 \pm 0.017}$
LSVM	0.202 ± 0.036	5	$\mathbf{0.178 \pm 0.037}$	4	0.188 ± 0.026
MLP	0.292 ± 0.077	4	0.194 ± 0.060	6	$\mathbf{0.186 \pm 0.021}$

TABLA 6.2. Error de clasificación en datos WPBC (con inf. adicional).

En la Tabla 6.2 se reportan los resultados obtenidos. Se observa que:

- El uso de información adicional W mejora los rendimientos de todos los clasificadores. En particular, en MLP hay una mejora relativa en media de $(0.292 - 0.194)/0.292 = 33.5\,\%$ para RKEF-AI y de $(0.292 - 0.186)/0.292 = 36.3\,\%$ para HTS, respecto del desempeño de las reducciones obtenidas sin utilizar la información adicional.

- Los rendimientos de RKEF-AI y HTS son similares, ya que alcanzan el valor óptimo en la misma cantidad de casos, con pequeñas diferencias. El mejor resultado ocurrió en LSVM, con ventaja relativa de RKEF-AI de $(0.188 - 0.178)/0.188 = 5.3\,\%$ respecto a HTS.

- La similitud en el rendimiento de RKEF-AI y HTS también se evidencia en las dimensiones $\hat{d}$ y $\hat{d}_{\text{env}}$ estimadas. En general, $\hat{d}_{\text{env}} \leq 6$ para ambos métodos.

De acuerdo a las observaciones, concluimos que usar la información adicional para este ejemplo fue provechoso. Además, la metodología en dos pasos evita la reducción brusca de dimensión $p = 30$ a $\hat{d} = 1$, estableciendo en el primer paso $3 \leq \hat{d}_{\text{env}} \leq 6$, lo cual permite pensar que dicho paso no solo tiene como tarea incorporar la información adicional, sino también mejorar de alguna manera las condiciones para la reducción final de $\mathbf{X}$.

A pesar de que en la comparación directa entre RKEF-AI y HTS hubo paridad, es importante remarcar que HTS tiene limitaciones que en este conjunto de datos no afectaron: efectúa reducciones lineales y se usó en un problema donde $\mathbf{X}$ es de baja dimensionalidad y W es univariada, discreta y con s pequeño.

		RKEF		RKEF-AI(COIR)		
Clasificador	Sin reducir	$\hat{d}$	Error	$\hat{d}$	$\hat{d}_{\text{env}}$	Error
LDA	0.085	18	0.077	20	400	**0.042**
QDA	0.689	28	0.059	20	400	**0.039**
LSVM	0.039	17	0.060	25	300	**0.031**

TABLA 6.3. Error de clasificación en datos PanCancer (con inf. adicional).

6.5.2 Datos de cáncer de páncreas

Analizaremos nuevamente el conjunto de datos PanCancer de la Subsección 5.2.2, pero añadiendo un conjunto de variables con información adicional. Es decir, estudiamos un problema de clasificación $Y|\mathbf{X}$ con $Y \in \{1, \ldots, 10\}$, $\mathbf{X} \in \mathbb{R}^{1000}$ y con información adicional $\mathbf{W} \in \mathbb{R}^{50}$.

Debido a las características de $\mathbf{W}$, para el método en dos pasos generalizado utilizaremos RKEF-AI(COIR); esto es, aplicaremos en primer lugar COIR para obtener $\mathbf{R}^{(1)}(\mathbf{X})$ y luego RKEF para la reducción final $\mathbf{R}(\mathbf{X}) = \mathbf{R}^{(2)}(\mathbf{R}^{(1)}(\mathbf{X}))$.

Las condiciones del problema no permiten aplicar HTS, por lo cual evaluaremos el método en dos pasos generalizado en contraste con la reducción vía RKEF sin información adicional, para ejemplificar la importancia de incluir $\mathbf{W}$ en el proceso.

Procedimiento *6.5.2. Datos PanCancer*

1° Obtener vía RKEF-AI(COIR) las reducciones $\hat{\mathbf{R}}_{\text{AI}}$, con $\hat{d}_{\text{env}} \in \{100, 150, \ldots, 500\}$ y $\hat{d} \in \{1, \ldots, 45\}$.

2° Obtener vía RKEF la reducción $\hat{\mathbf{R}}_{\text{SVM}}$ sin información adicional, de dimensión $\hat{d} \in \{1, \ldots, 45\}$.

3° Calcular el error de clasificación en los datos de prueba de los clasificadores LDA, QDA y LSVM. Utilizar *Criterio 2* para el parámetro de costo en LSVM.

En la Tabla 6.3 se reportan los resultados óptimos obtenidos, indicando para RKEF-AI(COIR) la dimensión $\hat{d}_{\text{env}}$ seleccionada. Se observa que:

- Usar la información adicional $\mathbf{W}$ mejora notablemente los resultados. Permitió lograr una mejora relativa de $(0.077 - 0.042)/0.077 = 45.5\,\%$ en LDA, $(0.059 - 0.039)/0.059 = 33.9\,\%$ en QDA y $(0.060 - 0.031)/0.060 = 48.3\,\%$ en LSVM.
- En LSVM, el empleo de información adicional permite conservar el rendimiento predictivo de los datos originales, corrigiendo así la pérdida de efectividad de RKEF.
- En todos los casos, el primer paso del método en dos pasos alcanzó una dimensión $\hat{d}_{\text{env}} \geq 300$. Esto permite tener un punto intermedio en la reducción brusca de $p = 1000$ a $\hat{d} \leq 30$.

De acuerdo a las observaciones, concluimos que nuevamente usar la información adicional mejoró las características de la reducción. Además, este ejemplo muestra el potencial del método en dos pasos generalizado cuando se combinan métodos de reducción adecuados tanto para la naturaleza de $\mathbf{X}$ como de la variable adicional $\mathbf{W}$. En particular, el uso de COIR para enfrentar el problema de la regresión de $(Y, \mathbf{W})$ en $\mathbf{X}$ permite seguir aprovechando el enfoque basado en núcleos, sin caer en un problema de dimensionalidad en RKEF. Esto permite condicionar mejor el problema a partir de la reducción intermedia del paso (I), para luego sí aprovechar la efectividad de RKEF en el paso (II), logrando finalmente una importante reducción de la dimensión de $\mathbf{X}$.

6.6 Comentarios de cierre de capítulo

En este capítulo desarrollamos una generalización del método en dos pasos de [Hung et al., 2015], una herramienta útil para afrontar situaciones en donde se cuenta con una variable adicional que contiene información acerca de la variable respuesta Y. Nuestra generalización permite dar solución a las limitaciones propias del modelo que se propone en dicho trabajo, abriendo la posibilidad de trabajar en escenarios de mayor complejidad.

En efecto, diseñamos una estrategia para incluir cualquier método de reducción, lo cual da al usuario la libertad de combinar el par de metodologías que considere más acordes al problema en cuestión. En particular, y para poner en evidencia el potencial de los métodos basados en núcleos, mostramos en un ejemplo con datos reales de alta dimensionalidad que la combinación COIR-RKEF puede ser muy acertada. En ambos métodos, los núcleos juegan un papel fundamental para extraer información pertinente sobre el objeto de estudio.

Conclusiones generales

En esta tesis presentamos nuevos métodos de aprendizaje estadístico destinados a reducir la dimensión de los datos en problemas de regresión y clasificación, sin perder información de la dependencia funcional entre los predictores y la respuesta. El atractivo principal de este tipo de métodos es ayudar a obtener resultados interpretables en el análisis de datos con características muy complejas y en los que la cantidad de variables predictoras puede exceder largamente la cantidad de ejemplos medidos.

Los métodos propuestos se han desarrollado en el marco del concepto de suficiencia estadística, el cual resulta clave para establecer un puente entre los objetivos de predicción y la necesidad de comprender los fenómenos que originan los datos. La literatura en torno a la reducción dimensional en el marco del aprendizaje automático es basta, pero durante mucho tiempo estuvo limitada esencialmente a precondicionar los datos para adecuarlos a la aplicación de métodos analíticos clásicos. No obstante, esos datos así reducidos ofrecen pocas garantías para muchas preguntas importantes: si no observamos en ellos diferencias globales entre grupos, ¿realmente no la hay?, ¿podemos usar la representación de baja dimensión para hacer inferencia sobre el problema original? La noción de suficiencia, por el contrario, nos ayuda a comprender relaciones globales entre predictores y respuesta, analizando la relación entre respuesta y reducciones suficientes de los predictores, un problema típicamente más fácil de modelar y que en ocasiones se puede también visualizar.

A diferencia de la mayor parte de los métodos existentes de reducción suficiente de dimensiones, los métodos propuestos implican transformaciones fuertemente no lineales de los predictores medidos. La primera contribución importante del trabajo es un método de reducción dimensional basado en una extensión infinito-dimensional de la familia

exponencial. Está motivado por numerosos resultados importantes disponibles para la familia exponencial que, sin embargo, muchas veces resultan difíciles de aplicar en problemas reales, debido a la dificultad de definir de antemano el modelo adecuado dentro de la familia. Luego, si bien la familia es amplia y presenta elementos capaces de describir una amplia variedad de datos con características muy diferentes, esa flexibilidad no se traslada a la práctica. Modelos diferentes requieren por lo general estimadores específicos y algoritmos especialmente adaptados, muchas veces complejos, sin contar la dificultad para verificar rigurosamente las suposiciones de modelado introducidas.

La reducción dimensional basada en KEF aprovecha de algún modo las ventajas del marco conceptual que provee la familia exponencial y, a la vez, evita muchos de sus problemas prácticos. La conexión con SVM nos ofrece una vía de estimación eficiente y unificada: la naturaleza específica de los datos puede afectar la elección del núcleo, pero no altera la maquinaria de aprendizaje. Más aún, un mismo núcleo pero con parámetros distintos ofrece soluciones con distinto grado de flexibilidad para tratar variaciones locales.

En esta tesis nos enfocamos en los aspectos conceptuales, pero hay más propiedades que pueden ser relevantes en aplicaciones. Por ejemplo, se puede acoplar la reducción dimensional basada en KEF a métodos de selección de variables basadas en núcleos, lo cual resulta en un núcleo compuesto que puede determinar la KEF a utilizar. Más aún, si ese núcleo se compone de la combinación lineal de núcleos elementales, la reducción resultante puede entenderse como superposición de reducciones. La estrategia puede utilizarse para visualizar la relevancia de diferentes grupos de variables sobre la respuesta. Por otra parte, la composición de núcleos puede facilitar el modelado de datos con características dispares, muchas veces difíciles de describir bajo un único modelo típico. Esta alternativa es común en el contexto de SVM, pero escasamente explorada en el caso de SDR.

En la segunda contribución importante de esta tesis, abordamos el problema de obtener provecho de información adicional disponible únicamente durante la etapa de entrenamiento, para eventualmente mejorar el desempeño de modelos predictivos. Este escenario no convencional formaliza algunas situaciones de interés práctico. Por ejemplo, en estudios clínicos, es común recolectar abundante cantidad de información que no es esperable

reproducir luego con pacientes en la práctica clínica habitual. No obstante, esa información adicional puede ser valiosa para el ajuste de herramientas de asistencia al diagnóstico o la evaluación predictiva de la respuesta de un paciente a determinadas intervenciones.

La estrategia presentada para aprovechar la información adicional consiste en buscar primero una transformación de los predictores comunes que preserve información sobre la respuesta de interés y también sobre la información adicional, para luego buscar una reducción posterior que se enfoque únicamente en la respuesta de interés. La cadena de transformaciones se aplica únicamente a los predictores comunes, disponibles tanto en la etapa de entrenamiento como en el uso estable de la herramienta aprendida. En este contexto, nuestro aporte posibilita el uso de información adicional de alta dimensión, mientras que las estrategias publicadas típicamente se limitan a casos de carácter categórico, o de muy baja dimensión en el caso de variables continuas. Por otra parte, las pruebas efectuadas muestran una mejora en la capacidad predictiva respecto al modelado que descarta la información adicional.

El trabajo efectuado podría extenderse en el futuro en varias direcciones. Por un lado, la familia exponencial finita presenta propiedades geométricas muy especiales que han motivado un área de estudio conocida como *geometría de la información*. Para el caso de KEF, algunas propiedades se estudiaron en [Fukumizu, 2009]. Dada la expresividad de esta extensión de la familia exponencial, es de interés explorar soluciones paramétricas al problema de *adaptación de dominio*, que ocurre cuando la distribución de los datos presenta cambios entre la etapa de entrenamiento de un modelo predictivo y su posterior uso práctico. Por otra parte, es de interés analizar si los resultados principales de esta tesis son extensibles a escenarios con condiciones más débiles que los RKHS. Por ejemplo, en ciertas aplicaciones es común contar con nociones de disimilaridad entre observaciones, propias de la disciplina, que no se traducen en núcleos definidos positivos. Algunas estrategias basadas en núcleos pueden extenderse a este escenario mediante adaptaciones de las herramientas analíticas originales [Oglic and Gärtner, 2018]. No obstante, la posibilidad de extender los resultados de esta tesis requiere un abordaje adecuado.

ANEXO **A**

Conceptos útiles

A.1 Métodos de clasificación

En lo que sigue, consideraremos un problema de clasificación binaria $Y|\mathbf{X}$, con $Y \in \{-1, 1\}$. Luego, cada método se extiende de forma natural a problemas de clasificación multiclase. Para un estudio más detallado, se sugiere revisar bibliografía específica [p. ej., Bishop, 2006; Hastie et al., 2009; Haykin, 2009].

A.1.1 Máquinas de vectores soporte (SVM) [Boser et al., 1992]

Sea $G : \mathbb{R}^p \to \{-1, 1\}$ una función de la forma $G(\mathbf{x}) = \text{sign}\,[f(\mathbf{x})]$, con $f : \mathbb{R}^p \to \mathbb{R}$ denominada *score*. La función G define un *clasificador* y la ecuación $f(\mathbf{x}) = 0$ determina en $\mathbb{R}^p$ una *frontera de decisión*. Veremos a continuación cómo, a partir de un conjunto de datos $\mathcal{D}_{(\mathbf{x},y)}$ de $(\mathbf{X}, Y)$, SVM formula un problema de optimización para hallar $\hat{G}$.

A.1.1.1. SVM lineal (LSVM)

Si $f(\mathbf{x}) = \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x} + \beta_0$, con $\boldsymbol{\beta} \in \mathbb{R}^p$ y $\beta_0 \in \mathbb{R}$, se define el problema de optimización

$$\begin{aligned} &\min_{\boldsymbol{\beta},\beta_0,\boldsymbol{\xi}} \frac{1}{2}\|\boldsymbol{\beta}\|_2^2 + C\sum_{i=1}^{n}\xi_i \\ &\text{sujeto a } \ \xi_i \geq 0,\ y_i f(\mathbf{x}_i) \geq 1 - \xi_i, \end{aligned} \tag{A.1}$$

donde las componentes de $\boldsymbol{\xi} \in \mathbb{R}^n$ se denominan *variables de holgura*. La solución $\hat{f}$ de (A.1) define, mediante $\hat{f}(\mathbf{x}) = 0$, un hiperplano separador que es óptimo en el sentido de que maximiza el margen de separación entre las clases, pero permitiendo que algunos datos no estén del lado correcto de dicho margen. Las variables de holgura penalizan dichos puntos y el parámetro C controla la cantidad de puntos penalizados. En la solución $\hat{f}(\mathbf{x}) = \hat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{x} + \hat{\beta}_0$, el vector $\boldsymbol{\beta}$ es de la forma

$$\hat{\boldsymbol{\beta}} = \sum_{i=1}^{n} \hat{\alpha}_i y_i \mathbf{x}_i,$$

con $\hat{\alpha}_i \neq 0$ para algunos datos, los cuales se denominan *puntos soporte*. Finalmente, el clasificador es

$$\hat{G}(\mathbf{x}) = \operatorname{sign}[\hat{f}(\mathbf{x})] = \operatorname{sign}\left[\hat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{x} + \hat{\beta}_0\right].$$

A.1.1.2. SVM no lineal

El método de SVM lineal puede extenderse a una versión no lineal a partir de los RKHS (ver Sección 2.1). La idea es mapear los datos $\mathbf{X}$ a un espacio característico $\mathcal{H}_{\mathcal{X}}$, posiblemente de dimensión infinita, mediante una transformación $\phi(\mathbf{X})$, y luego aplicar SVM lineal en el espacio característico a los datos transformados. Las fronteras de decisión lineales en $\mathcal{H}_{\mathcal{X}}$ se traducen generalmente en fronteras de decisión no lineales en el espacio original $\mathbb{R}^p$.

Sea $\mathcal{H}_{\mathcal{X}}$ un RKHS con núcleo reproductor $k : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}$ y sea $\phi : \mathbb{R}^p \to \mathcal{H}_{\mathcal{X}}$ el mapeo característico asociado; es decir, $\phi(\mathbf{x}) := k(\mathbf{x}, \cdot)$. Para todo $i, j = 1, \ldots, n$ se tiene que

$$\langle \phi(\mathbf{x}_i), \phi(\mathbf{x}_j) \rangle_{\mathcal{H}_{\mathcal{X}}} = k(\mathbf{x}_i, \mathbf{x}_j). \tag{A.2}$$

El estudio del problema (A.1) sobre los datos transformados $\{\phi(\mathbf{x}_1), \ldots, \phi(\mathbf{x}_n)\}$ deriva en el problema de optimización

$$\begin{aligned} &\min_{\beta, \beta_0, \boldsymbol{\xi}} \frac{1}{2}\|\beta\|^2_{\mathcal{H}_{\mathcal{X}}} + C\sum_{i=1}^{n} \xi_i \\ &\text{sujeto a } \ \xi_i \geq 0, \ y_i f(\mathbf{x}_i) \geq 1 - \xi_i, \end{aligned} \tag{A.3}$$

donde $\beta \in \mathcal{H}_{\mathcal{X}}$ y $f(\mathbf{x}) = \langle \beta, \phi(\mathbf{x}) \rangle_{\mathcal{H}_{\mathcal{X}}} + \beta_0$. La solución de (A.3) es de la forma

$$\hat{f}(\mathbf{x}) = \langle \hat{\beta}, \phi(\mathbf{x}) \rangle_{\mathcal{H}_{\mathcal{X}}} + \hat{\beta}_0, \tag{A.4}$$

con

$$\hat{\beta} = \sum_{i=1}^{n} \hat{\alpha}_i y_i \phi(\mathbf{x}_i). \tag{A.5}$$

Luego, el clasificador es

$$\hat{G}(\mathbf{x}) = \text{sign}[\hat{f}(\mathbf{x})] = \text{sign}\left[\langle \hat{\beta}, \phi(\mathbf{x}) \rangle_{\mathcal{H}_{\mathcal{X}}} + \hat{\beta}_0\right].$$

El *truco del núcleo* consiste en que únicamente necesitamos conocer k para evaluar $\hat{f}$ y clasificar un nuevo dato mediante $\hat{G}$. En efecto, reemplazando (A.5) en (A.4) y usando (A.2), resulta

$$\hat{f}(\mathbf{x}) = \sum_{i=1}^{n} \hat{\alpha}_i y_i \langle \phi(\mathbf{x}_i), \phi(\mathbf{x}) \rangle_{\mathcal{H}_{\mathcal{X}}} + \hat{\beta}_0 = \sum_{i=1}^{n} \hat{\alpha}_i y_i k(\mathbf{x}_i, \mathbf{x}) + \hat{\beta}_0.$$

A.1.1.3. SVM como método de regularización

SVM puede plantearse como un problema regularización en $\mathcal{H}_{\mathcal{X}}$ [Evgeniou et al., 1999] de la siguiente manera:

$$\min_{f \in \mathcal{H}_{\mathcal{X}}} \frac{1}{n} \sum_{i=1}^{n} (1 - y_i f(\mathbf{x}_i))_+ + \lambda \|f\|_{\mathcal{H}_{\mathcal{X}}}, \tag{A.6}$$

donde $[h(\mathbf{x})]_+ := \text{máx}\{0, h(\mathbf{x})\}$ denota la parte positiva de una función h. Si hacemos $n \to \infty$, podemos expresar (A.6) a nivel poblacional como

$$\min_{f \in \mathcal{H}_{\mathcal{X}}} \mathbb{E}\left[(1 - y_i f(\mathbf{x}_i))_+\right] + \lambda \|f\|_{\mathcal{H}_{\mathcal{X}}}. \tag{A.7}$$

En [Lin, 2002, Lema 3.1] se prueba que el minimizador de la función objetivo de (A.7) es la *regla de Bayes* dada por

$$h^*(\mathbf{x}) := \text{sign}\left[\mathbb{P}\{Y = 1 | \mathbf{X} = \mathbf{x}\} - \frac{1}{2}\right]. \tag{A.8}$$

A.1.2 Análisis discriminante lineal (LDA)

La regla de Bayes definida por (A.8) asigna $\mathbf{x}$ a la clase cuya probabilidad a posteriori $\mathbb{P}\{Y = y|\mathbf{X} = \mathbf{x}\}$ sea máxima. Así, $h^*(\mathbf{x}) = 1$ si y solo si

$$\mathbb{P}\{Y = 1|\mathbf{X} = \mathbf{x}\} > \mathbb{P}\{Y = -1|\mathbf{X} = \mathbf{x}\}. \tag{A.9}$$

Las probabilidades a posteriori en (A.9) pueden ser reformuladas, en virtud del Teorema de Bayes, de la siguiente manera:

$$\mathbb{P}\{Y = \pm 1|\mathbf{X} = \mathbf{x}\} = \frac{\mathbb{P}\{\mathbf{X} = \mathbf{x}|Y = \pm 1\} \cdot \mathbb{P}\{Y = \pm 1\}}{\mathbb{P}\{\mathbf{X} = \mathbf{x}\}}. \tag{A.10}$$

En (A.10) podemos considerar las funciones de densidad de las clases; para el caso binario escribimos $\mathbb{P}\{\mathbf{X} = \mathbf{x}|Y = \pm 1\} = p_{f^\pm}(\mathbf{x})$. Luego, reemplazando en (A.9), se obtiene

$$\pi_+ p_{f^+}(\mathbf{x}) > \pi_- p_{f^-}(\mathbf{x}), \tag{A.11}$$

donde $\pi_\pm := \mathbb{P}\{Y = \pm 1\}$. Así, modelando las densidades de cada clase, se puede deducir una fórmula de decisión a partir de la desigualdad (A.11). Bajo la suposición de que $\mathbf{X}|(Y = \pm 1) \sim \mathcal{N}_p(\boldsymbol{\mu}_\pm, \boldsymbol{\Delta})$, donde $\boldsymbol{\Delta}$ es común para todas las clases, se tiene

$$p_{f^\pm}(\mathbf{x}) = \frac{1}{(2\pi)^{n/2}|\boldsymbol{\Delta}|^{1/2}} \exp\left\{-\frac{1}{2}\langle \mathbf{x} - \boldsymbol{\mu}_\pm, \boldsymbol{\Delta}^{-1}(\mathbf{x} - \boldsymbol{\mu}_\pm)\rangle_{\mathbb{R}^p}\right\}. \tag{A.12}$$

Finalmente, de (A.11) y (A.12), el clasificador de Bayes resulta

$$h^*(\mathbf{x}) = \underset{y\in\pm 1}{\arg\max}\left\{\langle \mathbf{x}, \boldsymbol{\Delta}^{-1}\boldsymbol{\mu}_\pm\rangle_{\mathbb{R}^p} - \frac{1}{2}\langle \boldsymbol{\mu}_\pm, \boldsymbol{\Delta}^{-1}\boldsymbol{\mu}_\pm\rangle_{\mathbb{R}^p} + \log \pi_\pm\right\}. \tag{A.13}$$

El método *Análisis Discriminante Lineal* (LDA) surge de estimar (A.13) utilizando estimadores de $\boldsymbol{\mu}_\pm, \boldsymbol{\Delta}$ y $\pi_\pm$. Usando MLE, el clasificador LDA queda definido por

$$\hat{G}_{\mathrm{LDA}}(\mathbf{x}) := \underset{y\in\pm 1}{\arg\max}\left\{\langle \mathbf{x}, \hat{\boldsymbol{\Delta}}^{-1}\hat{\boldsymbol{\mu}}_\pm\rangle_{\mathbb{R}^p} - \frac{1}{2}\langle \hat{\boldsymbol{\mu}}_\pm, \hat{\boldsymbol{\Delta}}^{-1}\hat{\boldsymbol{\mu}}_\pm\rangle_{\mathbb{R}^p} + \log \hat{\pi}_\pm\right\}.$$

Desarrollando, resulta $\hat{G}_{\mathrm{LDA}}(\mathbf{x}) = 1$ si

$$\langle \mathbf{x}, \hat{\boldsymbol{\Delta}}^{-1}(\hat{\boldsymbol{\mu}}_+ - \hat{\boldsymbol{\mu}}_-)\rangle_{\mathbb{R}^p} > \frac{1}{2}\langle \hat{\boldsymbol{\mu}}_+ - \hat{\boldsymbol{\mu}}_-, \hat{\boldsymbol{\Delta}}^{-1}(\hat{\boldsymbol{\mu}}_+ - \hat{\boldsymbol{\mu}}_-)\rangle_{\mathbb{R}^p} - \log\frac{\hat{\pi}_+}{\hat{\pi}_-}.$$

A.1.3 Análisis discriminante cuadrático (QDA)

Una alternativa a LDA surge de suponer que cada clase tiene su propia matriz de covarianza, es decir considerar $\mathbf{X}(Y = \pm 1) \sim \mathcal{N}_p(\boldsymbol{\mu}_\pm, \boldsymbol{\Delta}_\pm)$. Un desarrollo análogo a LDA conduce al método *Análisis Discriminante Cuadrático* (QDA), cuyo clasificador es

$$\hat{G}_{\text{QDA}}(\mathbf{x}) := \underset{y \in \pm 1}{\arg\max} \Big\{ -\frac{1}{2}\langle \mathbf{x}, \hat{\boldsymbol{\Delta}}_\pm^{-1}\mathbf{x}\rangle_{\mathbb{R}^p} + \langle \mathbf{x}, \hat{\boldsymbol{\Delta}}_\pm^{-1}\hat{\boldsymbol{\mu}}_\pm\rangle_{\mathbb{R}^p} - \frac{1}{2}\langle \hat{\boldsymbol{\mu}}_\pm, \hat{\boldsymbol{\Delta}}_\pm^{-1}\hat{\boldsymbol{\mu}}_\pm\rangle_{\mathbb{R}^p} - \frac{1}{2}\log\left|\hat{\boldsymbol{\Delta}}_\pm\right| + \log\hat{\pi}_\pm \Big\}.$$

A.1.4 K vecinos más cercanos (KNN)

Para $K \in \mathbb{N}$, el método KNN consiste en estimar las probabilidades condicionales en función de los vecinos más cercanos, usando alguna medida de distancia entre las observaciones. Esto es,

$$\mathbb{P}\{Y = \pm 1 | \mathbf{x}\} \approx \frac{1}{K} \sum_{\mathbf{x}_i \in N_K(\mathbf{x})} \mathbf{1}(y_i = \pm 1),$$

donde $N_K(\mathbf{x})$ es el conjunto formado por las K observaciones más cercanas a $\mathbf{x}$. Luego, se aplica la regla de Bayes para clasificar a $\mathbf{x}$ a la clase con mayor probabilidad.

A.1.5 Perceptrón multicapa (MLP)

El *perceptrón* es un modelo matemático del funcionamiento de una neurona propuesto por Frank Rosenblatt (1928-1971) a mediados del siglo XX. Este modelo originó el desarrollo de una amplia gama de metodologías basada en redes neuronales. La idea básica es que el dato de entrada $\mathbf{x} \in \mathbb{R}^p$ es afectado por un vector de parámetros o pesos sinápticos $\mathbf{w} \in \mathbb{R}^p$ de la neurona, dando lugar a la salida $\langle \mathbf{x}, \mathbf{w}\rangle_{\mathbb{R}^p} + b$, donde $b \in \mathbb{R}$ es un parámetro extra. Luego, una función de activación f decidirá si la neurona se activa o no según el valor de $f(\langle \mathbf{x}, \mathbf{w}\rangle_{\mathbb{R}^p} + b)$.

El *Perceptrón Multicapa* (MLP) es una red neuronal formada por capas de neuronas, de manera tal que cada neurona de una capa se conecta a cada neurona de la siguiente capa. En dicha red se distinguen una capa de entrada, que está dedicada a ingresar los

valores de un dato $\mathbf{x}$, una o más capas ocultas, y una capa de salida que decide el valor de la respuesta Y. Por supuesto, cada neurona tiene asociado a ella un vector de parámetros y, en consecuencia, la cantidad de parámetros de un MLP depende de la cantidad de neuronas que lo componen.

A.2 Mediana heurística

Los métodos basados en núcleos extraen la información de una muestra $\{\mathbf{x}_i\}_{i=1}^n$ a partir de la matriz de Gram $\mathbf{K}$, cuyos elementos son $(\mathbf{K})_{ij} := k(\mathbf{x}_i, \mathbf{x}_j)$. Esto es lo que caracteriza el *truco del núcleo*, puesto que no se necesita conocer explícitamente el RKHS.

Ahora bien, si consideramos el núcleo Gaussiano (2.4), la matriz $\mathbf{K}$ dependerá de las distancias $\|\mathbf{x}_i - \mathbf{x}_j\|_2$ y del valor del parámetro σ. A la hora de seleccionar σ, su relación con dichas distancias es importante en el siguiente sentido [Garreau et al., 2017]:

- Si σ es lo suficientemente grande en relación a los valores de $\|\mathbf{x}_i - \mathbf{x}_j\|_2$, entonces los elementos de $\mathbf{K}$ tienden a uno.
- Si σ es lo suficientemente pequeño, los elementos de $\mathbf{K}$ tienden a cero.

En ambos casos, la información que $\mathbf{K}$ extrae sobre la distribución de $\mathbf{X}$ tiende a perderse por la mala elección de σ. Por esa razón, parece razonable elegir valores de σ que estén en escala con las distancias entre observaciones. En concordancia con esto, uno de los métodos más utilizados para determinar el ancho de banda σ deriva del cálculo de lo que se denomina *mediana heurística*, que está dada por

$$H_n := \text{median}\left\{\|\mathbf{x}_i - \mathbf{x}_j\|_2^2 : 1 \leq i < j \leq n\right\}. \tag{A.14}$$

Podemos decir que H_n es un buen representante del rango de valores que toman los valores $\|\mathbf{x}_i - \mathbf{x}_j\|_2^2$, aunque se puede reemplazar en (A.14) la mediana empírica por cualquier otro cuantil si se considerase conveniente. Finalmente, de la ecuación $\sigma^2 = H_n$ surge la expresión de σ basado en la mediana heurística:

$$\sigma_{\text{med}} := \sqrt{H_n}. \tag{A.15}$$

A.3 Resultado útil

Proposición A.1. *Sean $Y, W \in \mathbb{R}$ variables aleatorias. Si $W \sim N(\mu_W, \sigma_W^2)$ y $Y|W \sim N(aW + b, \sigma_Y^2)$, entonces $Y \sim N(a\mu_W + b, \sigma_Y^2 + a^2\sigma_W^2)$.*

Demostración. Las funciones características de $Y|W$ y W son

$$\phi_{Y|W}(t) := \mathbb{E}_Y\left[e^{itY}|W\right] = \exp\left\{i(aW+b)t - \frac{\sigma_Y^2 t^2}{2}\right\}$$

y

$$\phi_W(t) := \mathbb{E}\left[e^{itW}\right] = \exp\left\{i\mu_W t - \frac{\sigma_W^2 t^2}{2}\right\},$$

respectivamente. A partir de ellas, obtendremos la función característica de Y:

$$\begin{aligned}
\phi_Y(t) := \mathbb{E}\left[e^{itY}\right] &= \mathbb{E}_W\left[\mathbb{E}_Y\left[e^{itY}|W\right]\right] \\
&= \mathbb{E}_W\left[\exp\left\{i(aW+b)t - \frac{\sigma_Y^2 t^2}{2}\right\}\right] \\
&= \exp\left\{ibt - \frac{\sigma_Y^2 t^2}{2}\right\}\mathbb{E}\left[e^{iWat}\right] \\
&= \exp\left\{ibt - \frac{\sigma_Y^2 t^2}{2}\right\}\phi_W(at) \\
&= \exp\left\{ibt - \frac{\sigma_Y^2 t^2}{2}\right\}\exp\left\{i\mu_W at - \frac{\sigma_W^2 a^2 t^2}{2}\right\} \\
&= \exp\left\{i(a\mu_W + b)t - \frac{(\sigma_Y^2 + a^2\sigma_W^2)t^2}{2}\right\}.
\end{aligned}$$

Luego, dado que la función característica caracteriza las distribuciones, obtenemos que $Y \sim N(a\mu_W + b, \sigma_Y^2 + a^2\sigma_W^2)$, como queríamos probar. ■

ANEXO B

Demostraciones del Capítulo 3

B.1 Demostración del Corolario 3.4

Sea $p(\mathbf{x}|y)$ la función de densidad condicional de $\mathbf{X}|(Y = y)$. Por Teorema de Factorización (Teorema 3.3), existen funciones $g(\mathbf{t}, y)$ y $h(\mathbf{x})$ tales que

$$p(\mathbf{x}|y) = g(\mathbf{R}(\mathbf{x}), y)h(\mathbf{x}). \tag{B.1}$$

Sea $\mathbf{S}^{-1} : \mathbf{S}(\mathbb{R}^q) \to \mathbb{R}^q$ la función inversa de $\mathbf{S}$ restringida a su imagen $\mathbf{S}(\mathbb{R}^q)$. Entonces (B.1) puede reescribirse como

$$p(\mathbf{x}|y) = g((\mathbf{S}^{-1} \circ \mathbf{S} \circ \mathbf{R})(\mathbf{x}), y)h(\mathbf{x}).$$

Definiendo $\tilde{g}(\mathbf{t}, y) = g(\mathbf{S}^{-1}(\mathbf{t}), y)$, resulta

$$p(\mathbf{x}|y) = \tilde{g}((\mathbf{S} \circ \mathbf{R})(\mathbf{x}), y)h(\mathbf{x}),$$

y el Teorema de Factorización concluye la prueba. ■

ANEXO C

Demostraciones del Capítulo 4

C.1 Demostración del Teorema 4.1

Siguiendo la idea de [Cook, 2007], vamos a tratar a la respuesta Y como parámetro y probaremos que $\mathbf{R}(\mathbf{X})$ es un estadístico suficiente para Y. La función de densidad de $\mathbf{X}|(Y = y)$ dada por (4.1) puede reescribirse como

$$p(\mathbf{x}|y) = h(\mathbf{x})\frac{\exp\{f_y(\mathbf{x}) - \mathbb{E}\,[f]\,(\mathbf{x})\}}{Z(f_y)}, \tag{C.1}$$

donde $h := q_0 \exp\{\mathbb{E}\,[f]\}$. Por otro lado,

$$\sum_{y=1}^{r} \pi_y (f_y - \mathbb{E}\,[f]) = \sum_{y=1}^{r} \pi_y f_y - \mathbb{E}\,[f] = 0,$$

lo cual significa que el conjunto $\{f_y - \mathbb{E}\,[f] : y \in \mathcal{Y}\}$ es linealmente dependiente y

$$f_r - \mathbb{E}\,[f] = -\frac{1}{\pi_r} \sum_{y=1}^{r-1} \pi_y (f_y - \mathbb{E}\,[f]). \tag{C.2}$$

Definimos $G : \mathbb{R}^{r-1} \times \mathcal{Y} \to \mathbb{R}$ mediante $G(\mathbf{t}, y) = \exp\{g(\mathbf{t}, y)\}/Z(f_y)$, donde

$$g(\mathbf{t}, y) = \begin{cases} t_y & \text{si } y \in [r-1], \\ -\dfrac{1}{\pi_r} \displaystyle\sum_{i=1}^{r-1} \pi_i t_i & \text{si } y = r. \end{cases}$$

Luego, (C.1) puede factorizarse como $p(\mathbf{x}|y) = G(\mathbf{R}(\mathbf{x}), y)\, h(\mathbf{x})$ y, en consecuencia, aplicando el Teorema de Factorización (Teorema 3.3) se concluye la prueba. ■

C.2 Demostración de la Proposición 4.4

Sin pérdida de generalidad, consideremos

$$\mathbf{R}(\mathbf{X}) = (R_1(\mathbf{X}), \ldots, R_{r-1}(\mathbf{X})) \qquad \text{y} \qquad \mathbf{R}'(\mathbf{X}) = (R_1(\mathbf{X}), \ldots, R_{r-2}(\mathbf{X}), R_r(\mathbf{X})),$$

resultantes de excluir en el Teorema 4.1 las clases $y = r$ y $y = r-1$, respectivamente. De (C.2), para todo $\mathbf{x} \in \mathcal{X}$ podemos escribir

$$\begin{pmatrix} R_1(\mathbf{x}) \\ \vdots \\ R_{r-2}(\mathbf{x}) \\ R_r(\mathbf{x}) \end{pmatrix} = \begin{pmatrix} 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 0 \\ -\pi_1/\pi_r & -\pi_2/\pi_r & \cdots & -\pi_{r-2}/\pi_r & -\pi_{r-1}/\pi_r \end{pmatrix} \begin{pmatrix} R_1(\mathbf{x}) \\ \vdots \\ R_{r-2}(\mathbf{x}) \\ R_{r-1}(\mathbf{x}) \end{pmatrix}.$$

Esta es una transformación lineal inyectiva entre $\mathbf{R}$ y $\mathbf{R}'$. Luego, por Definición 3.5, ambas reducciones son equivalentes. ■

C.3 Demostración alternativa del Corolario 4.6

Sean $\mathbf{x}_1, \mathbf{x}_2 \in \mathcal{X}$. Observar que

$$\frac{p(\mathbf{x}_1|y)}{p(\mathbf{x}_2|y)} = \frac{q_0(\mathbf{x}_1)}{q_0(\mathbf{x}_2)} \exp\langle f_y, k(\mathbf{x}_1, \cdot) - k(\mathbf{x}_2, \cdot)\rangle_{\mathcal{H}_\mathcal{X}}$$

es constante respecto del valor de y si y solo si $\langle f_y, k(\mathbf{x}_1, \cdot) - k(\mathbf{x}_2, \cdot)\rangle_{\mathcal{H}_\mathcal{X}}$ lo es.

Supongamos $\langle f_y, k(\mathbf{x}_1, \cdot) - k(\mathbf{x}_2, \cdot)\rangle_{\mathcal{H}_\mathcal{X}} = C$, con C una constante respecto de y. Tomando esperanza respecto de la variable Y, se cumple

$$\langle f_y - \mathbb{E}[f], k(\mathbf{x}_1, \cdot) - k(\mathbf{x}_2, \cdot)\rangle_{\mathcal{H}_\mathcal{X}} = 0. \tag{C.3}$$

Ahora bien, dado que $\mathbb{E}[f] = \pi_1 f_1 + \pi_2 f_2$, desarrollando (C.3) para $y = 1$ se obtiene

$$\langle f_1 - f_2, k(\mathbf{x}_1, \cdot) - k(\mathbf{x}_2, \cdot)\rangle_{\mathcal{H}_\mathcal{X}} = 0$$

$$\langle f_1 - f_2, k(\mathbf{x}_1, \cdot)\rangle_{\mathcal{H}_\mathcal{X}} = \langle f_1 - f_2, k(\mathbf{x}_2, \cdot)\rangle_{\mathcal{H}_\mathcal{X}}$$

$$(f_1 - f_2)(\mathbf{x}_1) = (f_1 - f_2)(\mathbf{x}_2)$$

$$R(\mathbf{x}_1) = R(\mathbf{x}_2).$$

Luego, podemos afirmar que

$$\frac{p(\mathbf{x}_1|y)}{p(\mathbf{x}_2|y)} \text{ es constante como función de } y \Leftrightarrow R(\mathbf{x}_1) = R(\mathbf{x}_2),$$

y el Teorema de Lehmann Scheffé (Teorema 3.7) concluye la prueba. ■

C.4 Demostración del Teorema 4.8

Se tiene que

$$\begin{aligned}
f_1 - \mathbb{E}[f] &= \pi_2(f_1 - f_2) + \pi_3(f_1 - f_3) + \cdots + \pi_r(f_1 - f_r), \\
f_2 - \mathbb{E}[f] &= -\pi_1(f_1 - f_2) + \pi_3(f_2 - f_3) + \pi_4(f_2 - f_4) + \cdots + \pi_r(f_2 - f_r), \\
f_3 - \mathbb{E}[f] &= -\pi_1(f_1 - f_3) - \pi_2(f_2 - f_3) + \pi_4(f_3 - f_4) + \cdots + \pi_r(f_3 - f_r), \\
&\vdots \\
f_{r-1} - \mathbb{E}[f] &= -\pi_1(f_1 - f_{r-1}) - \cdots - \pi_{r-2}(f_{r-2} - f_{r-1}) + \pi_r(f_{r-1} - f_r)
\end{aligned}$$

Es decir, cada coordenada de la reducción $\mathbf{R}(\mathbf{X})$ de (4.3) puede escribirse como combinación lineal de las coordenadas de la reducción $\tilde{\mathbf{R}}(\mathbf{X})$ de (4.8). En consecuencia, existe una transformación $\mathbf{S} : \mathbb{R}^{r(r-1)/2} \to \mathbb{R}^{r-1}$ tal que

$$\mathbf{R}(\mathbf{X}) = (\mathbf{S} \circ \tilde{\mathbf{R}})(\mathbf{X}). \tag{C.4}$$

Por Teorema de Factorización (Teorema 3.3) para la SDR $\mathbf{R}(\mathbf{X})$ y usando (C.4), la función de densidad de $\mathbf{X}|(Y = y)$ puede factorizarse como

$$p(\mathbf{x}|y) = G((\mathbf{S} \circ \tilde{\mathbf{R}})(\mathbf{x}), y)\, h(\mathbf{x}) = \tilde{G}(\tilde{\mathbf{R}}(\mathbf{x}), y)\, h(\mathbf{x}), \tag{C.5}$$

con G tal como fue definida en la demostración del Teorema 4.1 (ver Apéndice C.1) y $\tilde{G}(\mathbf{t}, y) = G(\mathbf{S}(\mathbf{t}), y)$. Luego, de (C.5) y en virtud del Teorema de Factorización, se concluye la prueba. ■

C.5 Demostración del Lema 4.10

La función

$$g(\mathbf{x}) = \log \frac{\alpha(\mathbf{x})}{1-\alpha(\mathbf{x})}$$

es inyectiva como función de α. Además, de la expresión de $\mathcal{P}$ en (2.10) y utilizando el Teorema de Bayes, se deduce que

$$g(\mathbf{x}) = \log \frac{p_{f^+}(\mathbf{x})}{p_{f^-}(\mathbf{x})} + \text{cte} = f^+(\mathbf{x}) - f^-(\mathbf{x}) + \text{cte} = R(\mathbf{x}) + \text{cte}.$$

Por lo tanto, podemos afirmar que hay un mapeo uno a uno entre $\alpha(\mathbf{x})$ y $R(\mathbf{x})$.

Para concluir la prueba, basta mostrar que $f_q(\alpha)$ es inyectiva. Derivando la definición de f_q dada por (4.12) respecto de α, se tiene que

$$\begin{aligned}
f_q'(\alpha) &= \frac{1}{q-1} \frac{\left[\alpha^{\frac{1}{q-1}-1} + (1-\alpha)^{\frac{1}{q-1}-1}\right]\left[\alpha^{\frac{1}{q-1}} + (1-\alpha)^{\frac{1}{q-1}}\right] - \left[\alpha^{\frac{1}{q-1}} - (1-\alpha)^{\frac{1}{q-1}}\right]\left[\alpha^{\frac{1}{q-1}-1} - (1-\alpha)^{\frac{1}{q-1}-1}\right]}{\left[\alpha^{\frac{1}{q-1}} + (1-\alpha)^{\frac{1}{q-1}}\right]^2} \\
&= \frac{2}{q-1} \frac{\alpha^{\frac{1}{q-1}}(1-\alpha)^{\frac{1}{q-1}-1} + \alpha^{\frac{1}{q-1}-1}(1-\alpha)^{\frac{1}{q-1}}}{\left[\alpha^{\frac{1}{q-1}} + (1-\alpha)^{\frac{1}{q-1}}\right]^2} \\
&= \frac{2}{q-1} \frac{\alpha^{\frac{1}{q-1}-1}(1-\alpha)^{\frac{1}{q-1}-1}(\alpha + 1 - \alpha)}{\left[\alpha^{\frac{1}{q-1}} + (1-\alpha)^{\frac{1}{q-1}}\right]^2} \\
&= \frac{2}{q-1} \frac{[\alpha(1-\alpha)]^{\frac{1}{q-1}-1}}{\left[\alpha^{\frac{1}{q-1}} + (1-\alpha)^{\frac{1}{q-1}}\right]^2}.
\end{aligned}$$

Dado que $0 < \alpha < 1$ y $q > 1$, $f_q'(\alpha)$ es continua y positiva. Luego, $f_q(\alpha)$ es inyectiva. ■

C.6 Demostración del Teorema 4.11

Del Corolario 4.6 se sabe que $R(\mathbf{X}) = (f^+ - f^-)(\mathbf{X})$ es una SDR minimal de $\mathbf{X}$ para el problema de clasificación $Y|\mathbf{X}$. Entonces, por Lema 4.10 y Corolario 3.4, $f_q(\mathbf{X})$ también es una SDR de $\mathbf{X}$ para el problema de clasificación $Y|\mathbf{X}$. ■

C.7 Demostración del Teorema 4.12

Sean $1 \leq i < j \leq r$. Por Lema 4.10, existe un mapeo uno a uno entre $f_q^{i,j}(\mathbf{X})$ y la reducción $R_{i,j}(\mathbf{X}) = (f_i - f_j)(\mathbf{X})$, donde f_i y f_j son los parámetros funcionales de las densidades de $\mathbf{X} = (Y = i)$ y $\mathbf{X}|(Y = j)$ en $\mathcal{P}$, respectivamente.

Por lo tanto, existe un mapeo uno a uno entre $\mathbf{R}_{\text{SVM}}(\mathbf{X})$ y $\tilde{\mathbf{R}}(\mathbf{X})$ dada por (4.8), siendo esta última una SDR de $\mathbf{X}$ para el problema de clasificación $Y|\mathbf{X}$, en virtud del Teorema 4.8. Luego, basta aplicar el Corolario 3.4 para concluir la prueba. ■

ANEXO D

Demostraciones del Capítulo 6

D.1 Demostración de la Proposición 6.5

Si $\mathcal{B}$ es una base de $\mathcal{S}_{(Y,W)|\mathbf{Z}}$, entonces $(Y, W) \perp\!\!\!\perp \mathbf{Z}|\mathbf{P}_{\mathcal{B}}\mathbf{Z}$. En consecuencia, por [Cook, 1998, Proposición 4.6], resulta

$$Y \perp\!\!\!\perp \mathbf{Z}|\mathbf{P}_{\mathcal{B}}\mathbf{Z}.$$

Esto último significa que $\mathcal{S}_{(Y,W)|\mathbf{Z}}$ es un subespacio de reducción para la regresión de Y en $\mathbf{Z}$. Luego, como $\mathcal{S}_{Y|\mathbf{Z}}$ es el CS para la regresión de Y en $\mathbf{Z}$, por Definición 3.9 debe ser $\mathcal{S}_{Y|\mathbf{Z}} \subset \mathcal{S}_{(Y,W)|\mathbf{Z}}$. ■

D.2 Demostración de la Proposición 6.10

Si $\mathbf{R}(\mathbf{X})$ es una SDR de $\mathbf{X}$ para la regresión de $(Y, \mathbf{W})$ en $\mathbf{X}$, entonces $(Y, \mathbf{W}) \perp\!\!\!\perp \mathbf{X}|\mathbf{R}(\mathbf{X})$. Luego, por [Cook, 1998, Proposición 4.6], resulta

$$Y \perp\!\!\!\perp \mathbf{X}|\mathbf{R}(\mathbf{X}),$$

lo cual concluye la prueba. ■

D.3 Demostración del Teorema 6.11

La reducción $\mathbf{R}^{(1)}(\mathbf{X})$ verifica $(Y, \mathbf{W}) \perp\!\!\!\perp \mathbf{X}|\mathbf{R}^{(1)}(\mathbf{X})$. Entonces, por [Cook, 1998, Proposición 4.6], es cierto que

$$Y \perp\!\!\!\perp \mathbf{X}|\mathbf{R}^{(1)}(\mathbf{X}). \tag{D.1}$$

Dado que $\mathbf{R}(\mathbf{X})$ es función de $\mathbf{X}$, de (D.1) y por [Cook, 1998, Proposición 4.5], resulta

$$Y \perp\!\!\!\perp \mathbf{X}|(\mathbf{R}^{(1)}(\mathbf{X}), \mathbf{R}(\mathbf{X})). \tag{D.2}$$

Además, la reducción $\mathbf{R}(\mathbf{X})$ verifica

$$Y \perp\!\!\!\perp \mathbf{R}^{(1)}(\mathbf{X})|\mathbf{R}(\mathbf{X}). \tag{D.3}$$

Luego, de (D.2), (D.3) y por [Cook, 1998, Proposición 4.6], se concluye $Y \perp\!\!\!\perp \mathbf{X}|\mathbf{R}(\mathbf{X})$.

■

D.4 Demostración del Teorema 6.12

Por Teorema 4.1, $\mathbf{R}^{(1)}(\mathbf{X})$ es una SDR de $\mathbf{X}$ para el problema de clasificación $M|\mathbf{X}$ y $\mathbf{R}(\mathbf{X})$ es una SDR de $\mathbf{X}$ para el problema de clasificación $Y|\mathbf{R}^{(1)}(\mathbf{X})$. Pero $M|\mathbf{X}$ es equivalente al problema de clasificación $(Y, W)|\mathbf{X}$. En consecuencia, basta aplicar el Teorema 6.11 para concluir la prueba. ■

Bibliografía

Adragni, K. P. and Raim, A. M. (2014). ldr: An R software package for likelihood-based sufficient dimension reduction. *Journal of Statistical Software*, 61(3):1–21.

Aronszajn, N. (1943). La théorie des noyaux reproduisants et ses applications Première Partie. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 39, pages 133–153. Cambridge University Press.

Aronszajn, N. (1950). Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68(3):337–404.

Baker, C. R. (1973). Joint measures and cross-covariance operators. *Transactions of the American Mathematical Society*, 186:273–289.

Berlinet, A. and Thomas-Agnan, C. (2004). *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Springer Science & Business Media, New York, NY, USA.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer Science & Business Media, New York, NY, USA.

Boser, B. E., Guyon, I. M., and Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers. In *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, COLT '92, pages 144–152. Association for Computing Machinery.

Bura, E. and Cook, R. D. (2001). Estimating the structural dimension of regressions via parametric inverse regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(2):393–410.

Bura, E., Duarte, S., and Forzani, L. (2016). Sufficient reductions in regressions with exponential family inverse predictors. *Journal of the American Statistical Association*, 111(515):1313–1329.

Bura, E. and Forzani, L. (2015). Sufficient reductions in regressions with elliptically contoured inverse predictors. *Journal of the American Statistical Association*, 110(509):420–434.

Bura, E., Forzani, L., Arancibia, R. G., Llop, P., and Tomassi, D. (2022). Sufficient reductions in regression with mixed predictors. *Journal of Machine Learning Research*, 23(102):1–47.

Canu, S. and Smola, A. (2006). Kernel methods and the exponential family. *Neurocomputing*, 69(7–9):714–720.

Casella, G. and Berger, R. L. (2002). *Statistical Inference*. Duxbury Press, Pacific Grove, CA, USA, second edition.

Chiaromonte, F., Cook, R. D., and Li, B. (2002). Sufficient dimension reduction in regressions with categorical predictors. *The Annals of Statistics*, 30(2):475–497.

Cook, R. D. (1998). *Regression Graphics: Ideas for Studying Regressions Through Graphics*. John Wiley & Sons, New York, NY, USA.

Cook, R. D. (2007). Fisher lecture: Dimension reduction in regression. *Statistical Science*, 22(1):1–26.

Cook, R. D. and Forzani, L. (2008). Principal fitted components for dimension reduction in regression. *Statistical Science*, 23(4):485–501.

Cook, R. D. and Forzani, L. (2009). Likelihood-based sufficient dimension reduction. *Journal of the American Statistical Association*, 104(485):197–208.

Cook, R. D., Forzani, L. M., and Tomassi, D. R. (2011). LDR: A package for likelihood-based sufficient dimension reduction. *Journal of Statistical Software*, 39:1–20.

Cook, R. D. and Weisberg, S. (1991). Sliced inverse regression for dimension reduction: Comment. *Journal of the American Statistical Association*, 86(414):328–332.

de Jong, S. (1993). SIMPLS: An alternative approach to partial least squares regression. *Chemometrics and Intelligent Laboratory Systems*, 18(3):251–263.

Evgeniou, T., Pontil, M., and Poggio, T. (1999). A unified framework for regularization networks and support vector machines. *Technical Report, M.I.T. Artificial Intelligence Laboratory and Center for Biological and Computational Learning, Department of Brain and Cognitive Sciences*.

Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 222(594-604):309–368.

Fukumizu, K. (2009). Exponential manifold by reproducing kernel Hilbert spaces. *Algebraic and Geometric Methods in Statistics*, pages 291–306.

Fukumizu, K., Bach, F. R., and Jordan, M. I. (2009). Kernel dimension reduction in regression. *The Annals of Statistics*, 37(4):1871–1905.

Fukumizu, K. and Leng, C. (2014). Gradient-based kernel dimension reduction for regression. *Journal of the American Statistical Association*, 109(505):359–370.

Garreau, D., Jitkrittum, W., and Kanagawa, M. (2017). Large sample analysis of the median heuristic. *arXiv preprint arXiv:1707.07269*.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Science & Business Media, New York, NY, USA, second edition.

Haykin, S. (2009). *Neural Networks and Learning Machines*. Pearson Education, Upper Saddle River, NJ, USA, third edition.

Hung, H., Liu, C.-Y., and Horng-Shing Lu, H. (2015). Sufficient dimension reduction with additional information. *Biostatistics*, 17(3):405–421.

Hurvich, C. M. and Tsai, C.-L. (1989). Regression and time series model selection in small samples. *Biometrika*, 76(2):297–307.

Hyvärinen, A. (2005). Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6(24):695–709.

Ibañez, I., Forzani, L., and Tomassi, D. (2022). Generalized discriminant analysis via kernel exponential families. *Pattern Recognition*, 132:108933.

ICGC (2010). International network of cancer genome projects. *Nature*, 464(7291):993–998.

Kim, M. and Pavlovic, V. (2011). Central subspace dimensionality reduction using covariance operators. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(4):657–670.

Li, B., Artemiou, A., and Li, L. (2011). Principal support vector machines for linear and nonlinear sufficient dimension reduction. *The Annals of Statistics*, 39(6):3182–3210.

Li, B. and Wang, S. (2007). On directional regression for dimension reduction. *Journal of the American Statistical Association*, 102(479):997–1008.

Li, B., Zha, H., and Chiaromonte, F. (2005). Contour regression: A general approach to dimension reduction. *The Annals of Statistics*, 33(4):1580–1616.

Li, K.-C. (1991). Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association*, 86(414):316–327.

Li, K.-C. (1992). On principal Hessian directions for data visualization and dimension reduction: Another application of Stein's lemma. *Journal of the American Statistical Association*, 87(420):1025–1039.

Lin, Y. (2002). Support vector machines and the Bayes rule in classification. *Data Mining and Knowledge Discovery*, 6(3):259–275.

Mika, S., Rätsch, G., Weston, J., Schölkopf, B., and Müller, K.-R. (1999). Fisher discriminant analysis with kernels. In *Neural Networks for Signal Processing IX: Proceedings of the 1999 IEEE Signal Processing Society Workshop (Cat. No.98TH8468)*, pages 41–48. IEEE.

Naik, P. A. and Tsai, C.-L. (2005). Constrained inverse regression for incorporating prior information. *Journal of the American Statistical Association*, 100(469):204–211.

Nilsson, J., Sha, F., and Jordan, M. I. (2007). Regression on manifolds using kernel dimension reduction. In *Proceedings of the 24th International Conference on Machine Learning*, ICML '07, pages 697–704. Association for Computing Machinery.

Oglic, D. and Gärtner, T. (2018). Learning in reproducing kernel Kreın spaces. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 3859–3867. PMLR.

Roweis, S. T. and Saul, L. K. (2000). Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326.

Schölkopf, B., Smola, A., and Müller, K.-R. (1998). Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, 10(5):1299–1319.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464.

Shen, X.-J., Liu, S.-X., Bao, B.-K., Pan, C.-H., Zha, Z.-J., and Fan, J. (2020). A generalized least-squares approach regularized with graph embedding for dimensionality reduction. *Pattern Recognition*, 98:107023.

Sriperumbudur, B., Fukumizu, K., Gretton, A., Hyvärinen, A., and Kumar, R. (2017). Density estimation in infinite dimensional exponential families. *Journal of Machine Learning Research*, 18(57):1–59.

Sriperumbudur, B. K., Fukumizu, K., and Lanckriet, G. R. (2011). Universality, characteristic kernels and rkhs embedding of measures. *Journal of Machine Learning Research*, 12(70):2389–2410.

Steinwart, I. and Christmann, A. (2008). *Support Vector Machines*. Springer Science & Business Media, New York, NY, USA.

Steinwart, I., Hush, D., and Scovel, C. (2006). An explicit description of the reproducing kernel Hilbert spaces of Gaussian RBF kernels. *IEEE Transactions on Information Theory*, 52(10):4635–4643.

Tenenbaum, J. B., de Silva, V., and Langford, J. C. (2000). A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323.

Tomassi, D., Forzani, L., Duarte, S., and Pfeiffer, R. M. (2019). Sufficient dimension reduction for compositional data. *Biostatistics*, 22(4):687–705.

Wold, H. (1966). Estimation of principal components and related models by iterative least squares. *Multivariate Analysis*, pages 391–420.

Wold, H. (1975). Soft modeling by latent variables: The non-linear iterative partial least squares (NIPALS) approach. *Journal of Applied Probability*, 12(S1):117–142.

Wu, H.-M. (2008). Kernel sliced inverse regression with applications to classification. *Journal of Computational and Graphical Statistics*, 17(3):590–610.

Yan, S., Xu, D., Zhang, B., Zhang, H.-J., Yang, Q., and Lin, S. (2007). Graph embedding and extensions: A general framework for dimensionality reduction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(1):40–51.

Ye, Z. and Weiss, R. E. (2003). Using the bootstrap to select one of a new class of dimension reduction methods. *Journal of the American Statistical Association*, 98(464):968–979.

Yee, T. W. and Hastie, T. J. (2003). Reduced-rank vector generalized linear models. *Statistical Modelling*, 3(1):15–41.

Zhu, L., Wang, T., Zhu, L., and Ferré, L. (2010). Sufficient dimension reduction through discretization-expectation estimation. *Biometrika*, 97(2):295–304.

Doctorado en Ingeniería
mención Inteligencia Computacional, Señales y Sistemas

Título de la obra:

Nuevos métodos basados en núcleos para la representación eficiente de datos bajo suficiencia estadística

Autor: Diego Isaías Ibañez

Lugar: Santa Fe, Argentina

Palabras claves:

Análisis discriminante,
Reducción suficiente de dimensiones,
Espacios de Hilbert con núcleo reproductor,
Máquinas de vectores soporte,
Reducción suficiente de dimensiones con información adicional.