UNIVERSIDAD DE BUENOS AIRES
Facultad de Ciencias Exactas y Naturales
Departamento de Matemática

Selección de variables para datos multivariados y datos funcionales

Tesis presentada para optar al título de Doctor de la Universidad de Buenos Aires en el área Ciencias Matemáticas

Yanina Gimenez

Director de tesis: Dr. Ricardo Fraiman
Director asistente: Dra. Marcela Svarc
Consejero de estudios: Dra. Graciela Boente

Lugar de trabajo: Universidad de San Andrés

Fecha de Defensa: 27 de marzo de 2015

Selección de variables para datos multivariados y datos funcionales

Resumen

El problema de selección de variables es en la actualidad una de las principales áreas de investigación en la estadística. Si bien esta temática comenzó a analizarse en la década del 70, en los últimos años ha tenido un desarrollo explosivo, asociado a problemas de alta dimensión (high dimensional data) y de enormes bases de datos (big data). Estos desarrollos están vinculados fundamentalmente a los avances tecnológicos provenientes de problemas en biología, genética, meteorología, entre otras disciplinas.

En esta tesis trabajamos en el problema de selección de variables en diversos modelos estadísticos (regresión, clasificación, componentes principales, entre otros) para datos multivariados y para datos funcionales. Buscamos identificar un pequeño conjunto de variables que explique del mejor modo posible, mediante relaciones no paramétricas, el modelo en cuestión. Típicamente al analizar datos multivariados surgen dos tipos de problemáticas. Por un lado, encontramos variables no informativas, por otra parte, las variables suelen no ser independientes. El objetivo de esta tesis es entender la estructura interna de los datos asociados a cada modelo. Para realizarlo extendemos las ideas introducidas en Fraiman et al. (2008).

Primero damos una propuesta para seleccionar variables en el problema de componentes principales. Luego, introducimos una técnica general de selección de variables para datos multivariados. Estudiamos esta segunda propuesta para los modelos de regresión lineal, modelo lineal generalizado, componentes principales y correlación canónica. En todos los casos obtenemos resultados de consistencia. Mediante simulaciones describimos el comportamiento de los procedimientos presentados, realizamos comparaciones con otros métodos existentes e ilustramos con ejemplos de datos reales.

Finalmente extendemos las ideas del método propuesto a datos funcionales. En este caso no es razonable considerar las variables coordenadas como en el caso finito dimensional. Para ello, proponemos hacer la selección de un conjunto de funciones conocidas, $\{f_1, \ldots, f_p\}$ de las trayectorias, a valores reales. Dicho conjunto de funciones se seleccionan de acuerdo al problema a tratar. Hacemos diferentes propuestas de conjuntos que son adecuados para distintos problemas. El objetivo es seleccionar un pequeño subconjunto, $\{f_{i_1}, \ldots, f_{i_d}\}$, contenido en $\{f_1, \ldots, f_p\}$ que sea el que describa mejor el resultado del modelo estadístico aplicado. Damos una propuesta para los casos de clasificación, componentes principales y para el modelo lineal funcional con respuesta escalar y con respuesta funcional. En cada caso estudiamos resultados de consistencia.

Palabras Claves: Selección de Variables, Regresión, Componentes Principales, Clasificación, Datos Multivariados, Datos Funcionales.

Variable selection for multivariate data and functional data

Abstract

The study of variable selection problems in several statistical procedures is now a main–stream research area. These kind of problems have first been tackled in the 70's. As a result of the enormous technological advances it has become during the last years an enormous challenge, associated with high dimensional and big data problems. Biological, genetic, meteorological problems, among others, can be addressed from this perspective.

We herein introduce a general procedure for selecting variables, which can be applied to several classical multivariate and functional problems. We seek to identify a small subset of the original variables that can better explain, through nonparametric relationships, the model concerned. The method typically yields some noisy uninformative variables and some variables that are strongly related because of general dependence. The aim of this work is to help understand the underlying structures of a given data set. We extend the ideas introduced by Fraiman et al. (2008).

The thesis has three main chapters. First, we introduce a procedure for variable selection for principal components. Then, we study a general procedure for variable selection for multivariate data. We study these techniques for linear regression models, generalize linear models, principal components and canonical correlation. The asymptotic behavior of the proposed methods are analyzed. Simulations describing the behavior of the new procedures have been carried out and comparisons with several well known variable selection procedures had also been done. In addition, we also illustrate the performance of the procedure analyzing several real data examples.

Finally, we extend the ideas of the method to functional data framework. In this case it makes not sense to consider the coordinates of the variables, as in the finite dimensional case. Hence, we propose to select, from a set of known functions $\left\{f_1, \ldots, f_p, f_i : L^2[a,b] \to \mathbb{R}\right\}$, a subset of them. The group of function should be selected in relation with the statistical model. The final goal is to keep a subset of function $\{f_{i_1}, \ldots, f_{i_d}\}$ from $\left\{f_1, \ldots, f_p\right\}$ that better explain the model. We study the cases of classification, principal components and linear regression with scalar and functional response. In each case the asymptotic behavior of the proposed method has been studied.

Key Words: Variable Selection, Regression, Principal Component Analysis, Classification, Multivariate Data, Functional Data.

Agradecimientos

A Ricardo Fraiman, por su enorme ayuda que hizo posible la realización de esta Tesis. Por transmitir la matemática con tanta alegría y entusiasmo. Por su constante optimismo. Porque charlar sobre matemática con él es realmente placentero.

A Marce Svarc, por las tantas charlas, por todos sus consejos y su ayuda incondicional que hizo posible finalizar este trabajo. Y además, por brindarme su amistad.

A mis papas, Marisol y la Oma por su constante apoyo y cariño. Por su eterna comprensión. Y también a Mariano y Silvi por los momentos compartidos.

A Mauri, por su amor y su paciencia. Por hacerme feliz y siempre buscar como ayudarme. Y también a su familia por quererme y comprender mis ausencias en esta última etapa.

A las Monis, por ser mis amigas de la vida, por su continua amistad. Y a Flor, Caro y Deny por los buenos momentos compartidos.

A UdeSA y su gente, por ser mi lugar trabajo, por encontrar en ella tantas personas que me han acompañado estos años. En especial a mi grupo de amigos que hace divertido el ir a trabajar. A Lucas, Guido, la niña María, Mai, Belu, Rama, Allegri y Carlita por los almuerzos compartidos y las salidas, por hacerme reír en el día a día.

A Vero Moreno, por ser mi mejor amiga de la facu. Por las tantas charlas y buenos momentos. Y a todos mis otros amigos por hacer divertido el estudiar y las salidas compartidas, en especial a Mateo, Lucho, Adri, Paulita, Pato, Xime, Ani S., Lucas B., Vero Moyano, Lu Ch., Ale,...

A Nico y Hernan, por los viernes de pádel.

¡Muchísimas Gracias!

Índice general

Capítulo 1

Introducción

El problema de selección de variables es hoy en día una de las principales áreas de estudio de la estadística. Si bien estos problemas comenzaron a analizarse para los modelos lineales en la década del 70, fue recién en los últimos años, de la mano de los avances tecnológicos, que se volvió imperioso contar con buenas técnicas de selección de variables que se pudieran utilizar en un marco más general que el de regresión lineal. Los avances tecnológicos posibilitaron dos cosas. Por un lado, la recolección de datos se volvió más sencilla permitiendo contar con mucha información para analizar. Por ejemplo, los bancos tienen información sobre características económicas, hábitos de consumo y sociales de sus clientes y sus potenciales clientes. En los estudios médicos se analizan diversos parámetros de los pacientes. En sociología y ciencia política se realizan encuestas para entender características de la población. Como podemos notar, el almacenamiento de grandes de datos se popularizó. Por otra parte, la capacidad de cálculo también creció exponencialmente, dando lugar a que técnicas estadísticas computacionalmente costosas dejaran de serlo y produciendo que muchas de ellas (como, cluster, clasificación, modelos de regresión generalizado, componentes principales) cuya aplicación antes estaba relegada únicamente al uso de expertos, hoy sean herramientas disponibles para un grupo mucho más amplio de profesionales.

La combinación de estas dos situaciones dan lugar a la aparición de nuevos problemas. El hecho de que la recolección y el almacenamiento de datos sea accesible hace que en muchas ocasiones parte de la información recolectada sea irrelevante para el problema en cuestión y entonces al excluirlas se logra explicar el fenómeno de una manera más adecuada. Pero además predictores innecesarios añaden ruido a la estimación enmascarando las variables que realmente son importantes. A su vez el exceso de información genera colinealidad entre las variables produciendo estimaciones erróneas. Otro motivo por el cual se desea seleccionar variables informativas es un tema de costos, ya que si el modelo va a ser utilizado para la predicción podemos ahorrar tiempo y/o dinero al no medir predictores redundantes. Extraer la información relevante inherente al problema estadístico que se esté estudiando suele ser una tarea compleja y esta es la problemática que estudiaremos en esta tesis.

En la actualidad muchos procedimientos estadísticos buscan resolver estos problemas

cuando el número de variables estudiadas es mayor al número de observaciones, estos problemas aparecen típicamente ligados a genes o microarrays. Nuestro trabajo no contempla estas situaciones, es decir, nos limitamos a estudiar el caso clásico donde el número de observaciones es mayor que el número de variables.

Las técnicas de selección de variables son ad hoc al problema estadístico que se esté estudiando. En los últimos años se han desarrollado técnicas de selección de variables para diferentes tipos de problemas multivariados, entre ellos para el análisis de componentes principales, correlación canónica y clasificación supervisada y no supervisada.

El modelo estadístico más difundido y estudiado es el modelo de regresión lineal. Esto se debe a sus buenas propiedades teóricas, la facilidad de cálculo y que resulta un modelo útil y sencillo de interpretar en contextos muy diversos. Por este motivo es que muchos de los problemas que se estudian en estadística surgen en el modelo lineal y luego son analizados en otros contextos. El caso de selección de variables no escapa a esta generalidad.

En regresión lineal se considera un vector aleatorio $(\mathbf{X}, Y)$ donde $\mathbf{X} \in \mathbb{R}^p$, $Y \in \mathbb{R}$ y se busca describir mediante una transformación lineal de las variables explicativas $\mathbf{X}$ el comportamiento de la variable de respuesta Y. Para ser más precisos, consideramos el modelo de regresión lineal

$$Y = \mathbf{X}'\beta + \varepsilon, \tag{1.1}$$

donde $\varepsilon \in \mathbb{R}$ es el error del modelo, que es una variable aleatoria con esperanza 0 y $\beta \in \mathbb{R}^p$ es un vector de parámetros desconocido que representa la relación lineal entre $\mathbf{X}$ e Y. La forma más usual de hallar β es por mínimos cuadrados. Se busca el $\beta_0 \in \mathbb{R}^p$ tal que

$$\beta_0 = arg \min_{\beta} E\left((Y - \mathbf{X}'\beta)^2\right).$$

En la versión empírica se consideran $(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)$ vectores independientes e idénticamente distribuidos realizaciones del modelo (1.1). Llamamos β_n al estimador de mínimos cuadrados dado por

$$\beta_n = arg \min_{\beta} \sum_{j=1}^{n} \left(Y_j - \mathbf{X}_j'\beta\right)^2.$$

A continuación describimos brevemente algunos métodos que han sido propuestos para seleccionar variables en este modelo.

R^2 ajustado

Wherry (1931) introduce el R^2 ajustado, como medida de bondad de ajuste del modelo de regresión lineal (1.1). Un criterio clásico para seleccionar una cantidad fija de variables es elegir el subconjunto que maximice esta medida.

Dados $(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)$ vectores aleatorios independientes e idénticamente distribuidos que satisfacen el modelo (1.1), se define

$$R^2 = 1 - \frac{\sum_{j=1}^{n} \left(Y_j - \mathbf{X}_j'\beta_0\right)^2}{\sum_{i=1}^{n} \left(Y_j - \overline{Y}\right)^2},$$

donde $\overline{Y} = \sum_{j=1}^{n} Y_j/n$. A partir de esta medida se define el R^2 ajustado ($ADJR2$)

$$ADJR2 = R^2 - \frac{n-1}{n-d}\left(1 - R^2\right).$$

Observemos que el R^2 mide la proporción de la variabilidad de Y que es explicada por $\mathbf{X}$, y al definir $ADJR2$ se tiene en cuenta la dimensión en la que se está trabajando (d). Se puede notar que si decrece la dimensión aumenta el $ADJR2$, de esta forma al elegir el subconjunto que maximiza $ADJR2$ se está priorizando a los modelos más parsimoniosos.

AIC

El Criterio de Información de Akaike (AIC) fue introducido en Akaike (1974). Se considera el conjunto de modelos

$$Y = \mathbf{X}'\beta_d + \varepsilon, \tag{1.2}$$

donde β_d es un vector en $\mathbb{R}^p$ con d elementos no nulos, es decir, el modelo es función de d variables del vector $\mathbf{X}$ y $\varepsilon \in \mathbb{R}$ es el error del modelo con $E(\varepsilon) = 0$. Este criterio analiza la función de verosimilitud para cada uno de los modelos (1.2)

Dados $(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)$ vectores aleatorios independientes e idénticamente distribuidos que satisfacen el modelo (1.2), se tiene $f_{Y_j}\left(y; \mathbf{X}_j'\beta_d\right)$ función de densidad de Y_j que depende del modelo considerado, denotamos $\beta_{d,n}$ al estimador del modelo (1.2) y definimos

$$L_d(\beta_{d,n}) = ln\left(\prod_{j=1}^{n} f_{Y_j}\left(y; \mathbf{X}_j'\beta_{d,n}\right)\right), \tag{1.3}$$

el logaritmo de la función de verosimilitud del modelo con d parámetros.

El modelo seleccionado es aquel que minimice la función

$$AIC = -2L_d\left(\beta_{d,n}\right) + 2d, \tag{1.4}$$

que consta de dos términos, el primero decrece al mejorar el ajuste del modelo y el segundo es una penalidad que aumenta al crecer la dimensión.

BIC

Schwarz (1978) propone el Criterio de Información de Bayes (BIC) para seleccionar variables. Dados $(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)$ vectores aleatorios independientes e idénticamente distribuidos que satisfacen el modelo (1.2), se calcula

$$BIC = -2L_d\left(\beta_{d,n}\right) + ln(n)d, \tag{1.5}$$

donde $L_d\left(\beta_{d,n}\right)$ está dado por (1.3) y n es la cantidad de observaciones. El subconjunto seleccionado es aquel que minimiza (1.5).

Es claro que hay una estrecha similitud entre AIC (1.4) y BIC (1.5) ya que coinciden en el primero de sus dos términos. El término en común disminuye a medida que el ajuste mejora, mientras que el otro término es una penalidad en la dimensión del problema. Como $ln(n) > 2$ si $n \geq 8$, BIC tenderá a elegir modelos más parsimoniosos que AIC.

LASSO

Tibshirani (1996) introduce "Least Absolute Shrinkage and Selection Operator" (LASSO). Esta propuesta se diferencia de las anteriores ya que el parámetro no se estima por mínimos cuadrados o máxima verosimilitud sino que es una propuesta para estimar β que como subproducto suele ser adecuado para seleccionar variables.

Dados $(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)$ vectores aleatorios independientes e idénticamente distribuidos que satisfacen el modelo (1.1) el estimador de β está dado por

$$\beta_{lasso} = arg\min_{\beta} \sum_{j=1}^{n} \left(Y_j - \mathbf{X}_j'\beta\right)^2 \text{ sujeto a } \|\beta\|_1 = \sum_{i=1}^{p} |\beta_i| \leq t, \tag{1.6}$$

donde t es una constante. La restricción en norma L^1 sobre el vector de parámetros β suele verse reflejada en que algunos de los coeficientes se anulan y de este modo se puede entender como un criterio de selección de variables. Es claro que (1.6) se puede reescribir como

$$\beta_{lasso} = arg\min_{\beta} \sum_{j=1}^{n} \left(Y_j - \mathbf{X}_j'\beta\right)^2 + \lambda_1 \sum_{i=1}^{p} |\beta_i|,$$

con $\lambda_1 > 0$.

Efron et al. (2004) proponen el algoritmo "Least Angle Regression" (LARS) que sirve para implementar LASSO. Cuando el número de variables (p) es menor al número de observaciones (n) el costo del algoritmo es del orden $O\left(p^3 + np^2\right)$, que es el costo del estimador de mínimos cuadrados. Y el orden de convergencia es $O\left(n^3\right)$ cuando $p >> n$. Por la eficiencia del algoritmo y por los buenos resultados de LASSO, esta propuesta fue extendida exitosamente a otros modelos estadísticos.

Elastic Net

Zou & Hastie (2005) propusieron "Elastic Net" que es una generalización de LASSO. Primero introducen "Naïve Elastic Net", en este caso el estimador de β está dado por

$$\beta_{NEN} = arg\min_{\beta} \sum_{j=1}^{n} \left(Y_j - \mathbf{X}'_j\beta\right)^2 \text{ sujeto a } \|\beta\|_1 = \sum_{i=1}^{p} |\beta_i| \leq t_1 \text{ y } \|\beta\|_2^2 = \sum_{i=1}^{p} \beta_i^2 \leq t_2,$$

donde t_1 y t_2 con constantes. O de manera equivalente,

$$\beta_{NEN} = arg\min_{\beta} \sum_{j=1}^{n} \left(Y_j - \mathbf{X}'_j\beta\right)^2 + \lambda_1 \sum_{i=1}^{p} |\beta_i| + \lambda_2 \sum_{i=1}^{p} \beta_i^2,$$

con $\lambda_1 > 0$ y $\lambda_2 > 0$. La diferencia con LASSO (1.6) es que se le agrega una penalización en norma L^2 al vector β.

Sin embargo observan, mediante simulaciones, que el desempeño que tiene no es siempre el esperado y para solucionar el problema reescalan la solución convenientemente. El estimador propuesto es

$$\beta_{EN} = (1 + \lambda_2)\beta_{NEN}.$$

En el mismo trabajo proponen el algoritmo LARS–EN, basado en LARS. El orden de convergencia es el mismo que el de LARS en el caso en que el número de variables es menor que el de observaciones, en caso contrario proponen un algoritmo stepwise, que si se lo detiene luego de m pasos necesita $O\left(m^3 + pm^2\right)$ operaciones. Con simulaciones y ejemplos reales muestran que en pocos pasos alcanzan la solución deseada. Al comparar LASSO con Elastic Net se puede ver que la última propuesta es mejor cuando el número de variables es mayor que el número de observaciones.

SCAD y MCP

Fan & Li (2001) y Zhang (2010) proponen dos técnicas de selección de variables, "Smoothly Clipped Absolute Deviation Penalty" (SCAD) y "Minimax Concave Penalty" (MCP). Buscan generalizar LASSO proponiendo una función de penalización diferente a la propuesta por Tibshirani. El estimador del vector β de regresión lineal tanto en LASSO, en Elastic Net, en SCAD y MCP puede ser escrito como aquel que minimiza la función

$$Q_{\lambda,\gamma}(\beta) = \frac{1}{2n} \sum_{j=1}^{n} \left(Y_j - \mathbf{X}'_j\beta\right)^2 + \sum_{i=1}^{p} p_{\lambda,\gamma}\left(|\beta_i|\right),$$

donde el primer término representa a la función de pérdida clásica del método de mínimos cuadrados y el segundo término es una función de penalización.

En el caso de SCAD, está dada por

$$p_{\lambda,\gamma}(\theta) = \begin{cases} \lambda\theta & \text{si } 0 < \theta \leq \lambda, \\ \frac{\gamma\lambda\theta - 0.5(\theta^2+\lambda^2)}{\gamma-1} & \text{si } \lambda < \theta \leq \gamma\lambda, \\ \frac{\lambda^2(\gamma^2-1)}{2(\gamma-1)} & \text{si } \theta > \gamma\lambda, \end{cases}$$

con $\lambda \geq 0$ y $\gamma > 2$. Al derivarla obtenemos

$$p'_{\lambda,\gamma}(\theta) = \begin{cases} \lambda & \text{si } 0 < \theta \leq \lambda, \\ \frac{\gamma\lambda-\theta}{\gamma-1} & \text{si } \lambda < \theta \leq \gamma\lambda, \\ 0 & \text{si } \theta > \gamma\lambda. \end{cases}$$

Mientras que para MCP la función de penalización está dada por

$$p_{\lambda,\gamma}(\theta) = \begin{cases} \lambda\theta - \frac{\theta^2}{2\gamma} & \text{si } 0 < \theta \leq \gamma\lambda, \\ \frac{1}{2}\gamma\lambda^2 & \text{si } \theta > \gamma\lambda, \end{cases}$$

con $\lambda \geq 0$ y $\gamma > 1$ y al derivarla obtenemos

$$p'_{\lambda,\gamma}(\theta) = \begin{cases} \lambda - \frac{\theta}{\gamma} & \text{si } 0 < \theta \leq \gamma\lambda, \\ 0 & \text{si } \theta > \gamma\lambda. \end{cases}$$

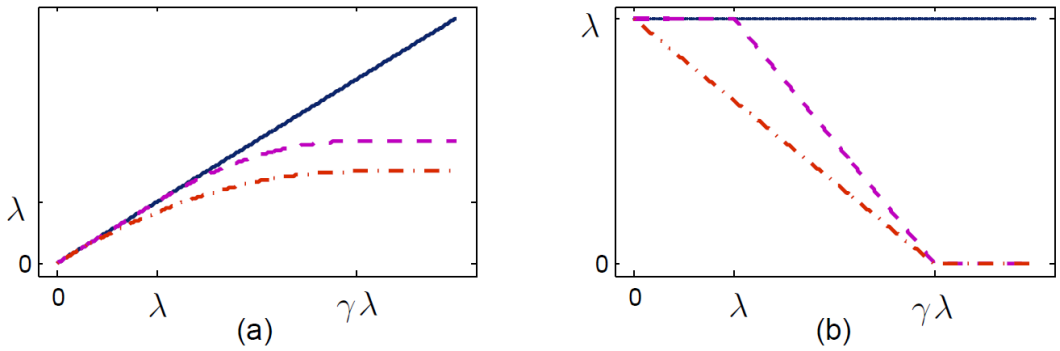

Figura 1.1: (a) Función de penalización. (b) Derivada de la función de penalización. Azul: LASSO. Violeta: SCAD. Rojo: MCP.

En la Figura 1.1(a) podemos observar las penalizaciones dadas a $|\beta_i|$ en los diferentes métodos. Se puede notar que SCAD y MCP para valores de $|\beta_i|$ grandes relajan la penalidad en cambio LASSO considera una función lineal. Esto se ve más claramente al mirar el gráfico de la derivada de la penalización (Figura 1.1(b)) donde se observa que para valores chicos de $|\beta_i|$ la penalización de SCAD y MCD se parece a la de LASSO pero mediante una función continua la van relajando hasta orden cero para valores de $|\beta_i| > \gamma\lambda$.

Estas propuestas son mucho más costosas computacionalmente que LASSO. Brehency & Huang (2011) proponen un algoritmo para implementar SCAD y MCP eficientemente.

Como vimos hasta aquí, la mayor parte de las propuestas de selección de variables buscan simultáneamente ajustar el modelo lineal y extraer las variables relevantes. Este patrón se repite cuando se analizan problemas de selección de variables en otros modelos estadísticos.

En esta tesis proponemos un enfoque diferente. En primer lugar realizamos el análisis estadístico en el espacio de dimensión alta, por ejemplo, regresión, componentes principales, clasificación, etc. Si consideramos que el análisis realizado fue exitoso entonces buscamos un pequeño subconjunto de variables que expliquen del mejor modo posible la salida del análisis original.

Para desarrollar estos métodos extendemos las ideas propuestas por Fraiman et al. (2008), donde se introducen dos propuestas para seleccionar variables en cluster y clasificación. Ambas propuestas se basan en la idea de cegar las variables innecesarias. Para cancelar los efectos de la variable sustituyen su valor por la esperanza (en la primera propuesta) y por la esperanza condicional (en la segunda). La primera propuesta tiene como objetivo identificar las variables ruidosas, mientras que la segunda también busca detectar problemas de dependencia. En este manuscrito generalizamos la idea de la segunda propuesta y la aplicamos en cada uno de los siguientes métodos: regresión, modelo lineal generalizado, componentes principales y correlación canónica.

Con el objetivo de hacer explícita la idea central de la tesis ilustramos con el siguiente ejemplo sencillo. Consideramos el modelo lineal dado por

$$Y = 3X_1 - 3X_2 + 4X_3 + \varepsilon, \tag{1.7}$$

donde $X_3 = exp(X_1 X_2)$ y X_1 y X_2 son variables aleatorias independientes, normales centradas en el origen, con varianza 1; el error, ε, es independiente, normalmente distribuido con esperanza cero y varianza 0.25. Generamos 100 observaciones con este modelo.

Ajustamos el modelo original y luego ajustamos el modelo estimando a la tercer variable con las otras dos, para lograrlo estimamos a la esperanza condicional $E(X_3|X_1, X_2)$ con el clásico criterio de vecinos más cercanos considerando 10 vecinos.

Se ve claramente que el modelo de regresión (1.7) es función de un subconjunto de cardinal dos de $\{X_1, X_2, X_3\}$, pero no existe un modelo lineal que este dado por 2 de las 3 variables que ajuste bien. En la Figura 1.2(a) observamos que son muy distintas las superficies al ajustar el modelo completo y el modelo considerando únicamente dos variables ($Y = \beta_1 X_1 + \beta_2 X_2 + \varepsilon$). Queda claro que necesitamos las tres variables para predecir correctamente a Y. En la Figura 1.2(b) vemos el scatter plot de $\mathbf{X}'\beta_n$ versus $\hat{\mathbf{X}}'\beta_n = \left(X_1, X_2, \hat{E}\left(X_3|X_1, X_2\right)\right)' \beta_n$, donde β_n es el estimador de mínimos cuadrados de β para el modelo completo. La relación lineal que se puede apreciar en esta figura muestra que predijimos muy bien a X_3. Siendo de este modo evidente que casi toda la información contenida en estas variables puede ser descrita por dos de ellas al estimar la tercera mediante la esperanza condicional.

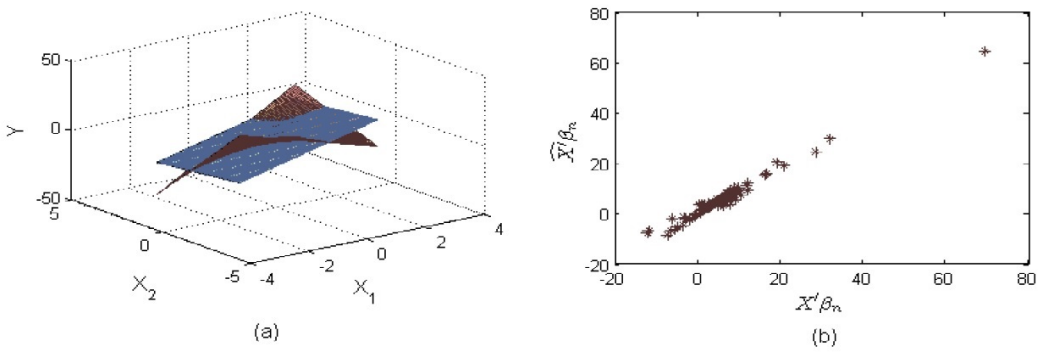

Figura 1.2: (a) Superficie púrpura: Ajuste usando el modelo lineal que involucra a las tres variables. Superficie azul: Ajuste usando un modelo lineal que solo involucra a las variables X_1 y X_2. (b) Eje horizontal: Función de regresión del conjunto de datos originales. Eje vertical: Función de regresión cuando X_3 es predicha por X_1 y X_2.

Esta es la idea central de esta tesis y la desarrollamos especialmente en el Capítulo 4 donde será aplicada a diferentes modelos estadísticos.

Lo que resta de la tesis se estructura del siguiente modo. En el Capítulo 2 damos las notaciones y definiciones preliminares para el desarrollo de la tesis. En el Capítulo 3 introducimos una propuesta para seleccionar variables en el problema de componentes principales. La propuesta general para abordar el problema de seleccionar variables en el contexto de regresión y reducción de la dimensión, la detallamos en el Capítulo 4. En el Capítulo 5 extendemos las ideas trabajadas al caso de datos funcionales. Finalmente, en el último Capítulo se encuentran las demostraciones de los teoremas enunciados.

Capítulo 2

Definiciones y Resultados preliminares

2.1 El Modelo de Regresión

El objetivo del análisis de regresión es comprender como una variable de respuesta Y ($Y \in \mathbb{R}$) está relacionada con un vector de variables $\mathbf{X} \in \mathbb{R}^p$. La forma más simple de vincularse es el modelo de regresión lineal (1.1).

El modelo de regresión se define del siguiente modo,

$$Y = g(\mathbf{X}, \beta) + \varepsilon, \tag{2.1}$$

donde $\varepsilon \in \mathbb{R}$ es la variable del error del ajuste con $E(\varepsilon) = 0$ y $\beta \in \mathbb{R}^p$ es el vector de parámetros desconocido a ser estimado. Si $g(\mathbf{X}, \beta) = \mathbf{X}'\beta$ estamos en el caso clásico, el modelo lineal (1.1), en caso contrario es un modelo no lineal.

El objetivo es hallar el vector $\beta_0 \in \mathbb{R}^p$,

$$\beta_0 = arg \min_{\beta} E\left(\rho\left(Y - g(\mathbf{X}, \beta)\right)\right),$$

donde $\rho(\cdot)$ es una función de pérdida. Si $\rho(\cdot) = (\cdot)^2$ obtenemos el estimador de mínimos cuadrados. En presencia de outliers se consideran funciones que den menor peso a residuos "grandes", entre ellas encontramos la familia de funciones de la bicuadrada de Tukey, que es redescendiente. Y está definida por

$$\rho(u) = \begin{cases} \frac{u^2}{2}\left(1 - \frac{u^2}{c^2} + \frac{u^4}{3c^4}\right) & \text{si } |u| \leq c, \\ \frac{c^2}{6} & \text{si } |u| > c. \end{cases}$$

2.2 Modelo Lineal Generalizado

Los modelos de regresión lineal dejan de ser apropiados cuando la variable de respuesta Y es cualitativa, por ejemplo binaria. Para resolver el problema, se utilizan los modelos lineales

generalizados (GLIM), introducidos por Nelder y Wedderburn (ver por ejemplo McCullagh & Nelder (1989) para un estudio detallado). Con este enfoque se logra unificar a varios modelos, incluyendo la regresión lineal, la logística y la regresión de Poisson entre otros. La solución es sencilla e ingeniosa. En un modelo GLIM se supone que el valor esperado, μ, de la variable de respuesta Y depende de las variables explicativas $\mathbf{X}$ a través de un vínculo g (función de link) no necesariamente lineal,

$$E(Y) = \mu = g^{-1}(\mathbf{X}'\beta). \tag{2.2}$$

La función de link es una transformación del parámetro de la distribución de Y a $\mathbb{R}$. Por ejemplo, si la distribución de Y es $Binomial(1, p)$ las funciones de vínculo g más habituales son $ln(p/(1-p))$ (modelo logit) o $\Phi^{-1}(p)$ (modelo probit), donde Φ es la función de distribución acumulada de una $Normal(0, 1)$.

Se asume que la distribución de la variable aleatoria Y pertenece a una familia exponencial, es decir su función de densidad f_Y puede expresarse de la forma

$$f_Y(y, p) = h(y)exp(\xi(p)T(y) - A(p)),$$

donde $h(y)$, $\xi(p)$, $T(y)$ y $A(p)$ son funciones conocidas.

Siendo entonces las componentes del modelo

- la familia exponencial de distribuciones f_Y,
- un predictor lineal $\eta = \mathbf{X}'\beta$,
- una función de link g para la que se verifica (2.2).

Para hallar $\beta_0 \in \mathbb{R}^p$ en general se considera el estimador de máxima verosimilitud.

2.3 Componentes Principales

El análisis de componentes principales es una técnica clásica de reducción de dimensión. El objetivo es transformar un conjunto de datos p-dimensional en otro de menor dimensión con la menor pérdida de información posible. Buscando tener, por ejemplo, una representación gráfica del conjunto en dos o tres dimensiones. El análisis de componentes principales busca direcciones ortogonales en las cuales proyectar los datos. La primera componente está dada por la dirección que maximiza la varianza de los datos proyectados. La segunda componente se busca, siguiendo el mismo criterio, en el espacio ortogonal a la primera componente, y así sucesivamente.

Más formalmente, sea $\mathbf{X} \in \mathbb{R}^p$ un vector aleatorio, el objetivo es hallar la dirección en la cual se maximiza la varianza de la proyección unidimensional $\alpha'\mathbf{X}$, es decir, hallar $\alpha_1 \in \mathbb{R}^p$ tal que,

$$\alpha_1 = arg \max_{\|\alpha\|=1} Var(\alpha'\mathbf{X}) = arg \max_{\|\alpha\|=1} \alpha'\mathbf{\Sigma}\alpha,$$

donde $\mathbf{\Sigma}$ es la matriz de covarianza de $\mathbf{X}$. Así obtenemos la primera componente principal que es el vector α_1 premultiplicado por el vector $\mathbf{X}$.

Asumimos que $\mathbf{\Sigma}$ es definida positiva y que todos sus autovalores $\lambda_1 > \cdots > \lambda_p$ son distintos. Los pesos de las siguientes componentes principales están definidos como

$$\begin{aligned} \alpha_k &= arg\max Var\,(\alpha'\mathbf{X}) = arg\max \alpha'\mathbf{\Sigma}\alpha \\ &\quad \text{sujeto a } \|\alpha\| = 1 \text{ y } <\alpha, \alpha_j> = 0 \text{ para } j = 1, \ldots, k-1, \end{aligned} \tag{2.3}$$

para $k = 2, \ldots, p$, donde $< \cdot, \cdot >$ es el producto interno usual en $\mathbb{R}^p$.

Del Teorema Espectral se concluye que si $\lambda_1 > \lambda_2 > \ldots > \lambda_p$, son los autovalores de $\mathbf{\Sigma}$, entonces las componentes principales se obtienen a partir de sus correspondientes autovectores α_k, para $k = 1, \ldots, p$. Luego, la k-ésima componente principal está determinada por

$$U_k = \alpha_k'\mathbf{X}. \tag{2.4}$$

Los autovalores asociados son una medida de la cantidad de información explicada por las componentes principales, más precisamente la k-ésima componente principal explica un $\left(\lambda_k / \sum_{i=1}^{p} \lambda_i\right)$ % de la variabilidad total.

En presencia de *outliers* se pueden obtener componentes principales robustas considerando una matriz robusta de correlación o bien siguiendo las indicaciones de Maronna et al. (2006).

Observación 1 *Hay ocasiones donde se considera que hay efectos distorsivos generados por la presencia de variables representadas en distintas unidades de medida, es preferible calcular las componentes principales utilizando la matriz de correlación en lugar de la matriz de covarianza.*

2.4 Correlación Canónica

El objetivo del análisis de correlación canónica es buscar las relaciones entre dos conjuntos de variables. El modo en que se realiza el análisis es similar al de componentes principales.

Se tienen dos vectores aleatorios $\mathbf{X} \in \mathbb{R}^p$ e $\mathbf{Y} \in \mathbb{R}^q$ y el objetivo es obtener una proyección lineal de cada uno de ellos que maximice la asociación entre los vectores originales.

Se busca $\alpha \in \mathbb{R}^p$ y $\beta \in \mathbb{R}^q$ tales que maximicen la correlación entre $\alpha'\mathbf{X}$ y $\beta'\mathbf{Y}$, es decir

$$\begin{aligned} &\text{maximizar } \alpha'\mathbf{\Sigma_{XY}}\beta \\ &\text{sujeto a } \alpha'\mathbf{\Sigma_X}\alpha = 1, \beta'\mathbf{\Sigma_Y}\beta = 1, \end{aligned}$$

donde $\mathbf{\Sigma_{XY}}$ es la matriz de covarianza entre $\mathbf{X}$ e $\mathbf{Y}$ y $\mathbf{\Sigma_X}$ y $\Sigma_\mathbf{Y}$ son las matrices de covarianza de $\mathbf{X}$ e $\mathbf{Y}$ respectivamente. α_1 y β_1 son los pesos correspondientes a la primera correlación canónica. Imponiendo además condiciones de ortogonalidad, de manera análoga al caso de

componentes principales, se obtienen los pesos correspondientes a las siguientes correlaciones canónicas. $\alpha_k \in \mathbb{R}^p$ y $\beta_k \in \mathbb{R}^q$ son los vectores que resuelven el siguiente problema de optimización

$$\begin{aligned} &\text{maximizar } \alpha' \mathbf{\Sigma_{XY}} \beta \\ &\text{sujeto a } \alpha' \mathbf{\Sigma_X} \alpha = 1,\ \beta' \mathbf{\Sigma_Y} \beta = 1, \\ &\alpha_j' \mathbf{\Sigma_X} \alpha = 0,\ \beta_j' \mathbf{\Sigma_Y} \beta = 0 \text{ para } j = 1, \ldots, k-1, \end{aligned} \tag{2.5}$$

para $k = 2, \ldots, \widetilde{p}$, donde $\widetilde{p}$ es el rango de la matriz $\mathbf{\Sigma_{XY}}$.

Sea

$$\kappa = \mathbf{\Sigma_X}^{-1/2} \mathbf{\Sigma_{XY}} \mathbf{\Sigma_Y}^{-1/2}.$$

Si γ_k y δ_k son los autovectores estandarizados de $\kappa\kappa'$ y $\kappa'\kappa$ respectivamente, las combinaciones lineales que nos determinan las variables canónicas son $\alpha_k = \mathbf{\Sigma_X}^{-1/2} \gamma_k$ y $\beta_k = \mathbf{\Sigma_Y}^{-1/2} \delta_k$, para $k = 1, \ldots, \widetilde{p}$. Luego, las k-ésimas variables canónicas son,

$$\begin{aligned} U_k &= \alpha_k' \mathbf{X}, \\ V_k &= \beta_k' \mathbf{Y}. \end{aligned} \tag{2.6}$$

A la correlación entre U_k y V_k se la denomina la k-ésima correlación canónica.

2.5 Definiciones y Notación

A continuación establecemos la notación que utilizamos a lo largo de la tesis.

Sea $\mathbf{X} \sim P \in \mathcal{P}_0$ un vector aleatorio en $\mathbb{R}^p$, donde $\mathcal{P}_0$ es un conjunto de distribuciones en probabilidad sobre $\mathbb{R}^P$. Sea $\mathbb{G}$ un conjunto de acciones y $\mathbb{E}$ un espacio métrico. Un *modelo estadístico poblacional* está dado por una función $\psi(P) := \psi(\mathbf{X}, g)$, donde $\psi : \mathbb{R}^p \times \mathbb{G} \to \mathbb{E}$. La salida del procedimiento estadístico es un elemento aleatorio en $\mathbb{E}$ dado por $\psi(\mathbf{X}, g)$.

A las coordenadas del vector $\mathbf{X}$ las denotamos $X[i]$, $i = 1, \ldots, p$.

Dado un conjunto de índices $I \subset \{1, \ldots, p\}$ con cardinal $d \leq p$, llamamos $\mathbf{X}(I)$ al conjunto de variables aleatorias $\{X[i], i \in I\}$. Más aún, haciendo abuso de notación, si $I = \{i_1 < \ldots < i_d\}$, notamos por $\mathbf{X}(I)$ al vector $(X[i_1], \ldots, X[i_d])$.

Definición 1 *Sea I un subconjunto de $\{1, \ldots, p\}$ llamamos **vector blinded** de $\mathbf{X}$ respecto de las coordenadas I al vector $\mathbf{Z}(I) \in \mathbb{R}^p$, tal que*

$$Z(I)[i] = \begin{cases} X[i] & \text{si } i \in I, \\ E\left(X[i] | \mathbf{X}(I)\right) & \text{si } i \notin I. \end{cases} \tag{2.7}$$

Notamos $Q(I)$ a la distribución de $\mathbf{Z}(I)$.

Observación 2 *En presencia de outliers es más adecuado considerar un procedimiento robusto en la definición del vector blinded. En (2.7) se puede considerar la mediana condicional, med*$(X[i]|\boldsymbol{X}(I))$ *en lugar de la esperanza condicional,* $E(X[i]|\boldsymbol{X}(I))$.

Notamos $\mathcal{I}_d$ a la familia de todos los conjuntos de $\{1, \ldots, p\}$ con cardinal d con $1 \leq d < p$.

Denotamos con $\|\cdot\|$ a la norma Euclídea y con $<\cdot,\cdot>$ al producto interno en $\mathbb{R}^p$. En caso contrario aclaramos en que espacio estamos considerando la norma y/o el producto interno.

Sea A un conjunto, notamos I_A a la función indicadora, es decir,

$$I_A(x) = \begin{cases} 1 & \text{si } x \in A, \\ 0 & \text{si } x \notin A. \end{cases}$$

2.6 Selección de Variables para Cluster y Clasificación

Fraiman et al. (2008) proponen dos técnicas de selección de variables para los problemas de aprendizaje supervisado y no supervisado. La primera tiene como objetivo identificar las variables ruidosas, mientras que la segunda también busca detectar el problema de dependencia. Ambas propuestas fueron concebidas para ser utilizadas luego de analizar los datos en el espacio original y estar conformes con los resultados obtenidos. Buscan un pequeño subconjunto de las variables originales que expliquen del mejor modo posible la salida del análisis original. Para realizarlo definen una función $h(\cdot)$ que mide la cercanía entre el procedimiento considerando las variables originales y el procedimiento al cegar las variables no seleccionadas. Para cancelar los efectos del conjunto de variables no elegidas, sustituyen su valor por la esperanza en la primera propuesta y por la esperanza condicional en la segunda.

Más formalmente, sea $\mathbf{X}' = (X[1], \ldots, X[p])$ un vector aleatorio en $\mathbb{R}^p$ con distribución P y g un procedimiento estadístico que da una partición del espacio. Para un número fijo de K grupos, consideran $g : \mathbb{R}^p \to \{1, \ldots, K\}$ una función que asigna a cada punto de $\Omega \subseteq \mathbb{R}^p$ a un único grupo. Por lo tanto, $G_k = g^{-1}(k)$ con $k = 1, \ldots, K$ es una partición disjunta de Ω. Luego de obtener una partición del espacio con la cual se está conforme, se busca un pequeño conjunto $d < p$ de variables que mantenga la partición del mejor modo posible. Para encontrar dicho subconjunto de variables definen para cada conjunto I con cardinal d, $\mathbf{Z}(I) \in \mathbb{R}^p$ tal que

$$Z(I)[i] = \begin{cases} X[i] & \text{si } i \in I, \\ E\left(X[i]\right) & \text{si } i \notin I, \end{cases}$$

para la primera propuesta y $\mathbf{Z}(I)$ como en (2.7) para la segunda. Buscan un conjunto I tal que $\psi(\mathbf{X}, g)$ esté lo más cerca posible de $\psi(\mathbf{Z}(I), g)$. Donde la idea de cercanía depende del proceso que se usa para analizar los datos y está determinada por la función $h(I, P, Q(I), \psi) := h(I)$. Ellos definen la función objetivo,

$$h(I) = \sum_{k=1}^{K} P\left(g(\mathbf{X}) = k, g(\mathbf{Z}(I)) = k\right),$$

que mide cuan semejantes son las particiones del espacio considerando el vector $\mathbf{X}$ y el vector $\mathbf{Z}(I)$. Finalmente buscan la familia de conjuntos $\mathcal{I}_0 \subset \mathcal{I}_d$ que verifiquen

$$\mathcal{I}_0 = arg \max_{I \in \mathcal{I}_d} h(I),$$

y eligen un conjunto de variables que pertenezca a $\mathcal{I}_0$.

Capítulo 3

Selección de Variables en Componentes Principales

En este capítulo damos una primera propuesta para extender la idea introducida por Fraiman et al. (2008) al modelo de componentes principales. Obtenemos resultados de consistencia, damos consideraciones prácticas para su implementación e ilustramos con un ejemplo con datos reales.

3.1 Introducción

El análisis de componentes principales es un método clásico de reducción de dimensionalidad. Su principal desventaja es que como las nuevas coordenadas son combinaciones lineales de las originales en muchos casos son difíciles de interpretar. Por este motivo, es relevante el problema de selección de variables en este contexto, ya que esto ayuda a mejorar la interpretación de los datos.

A continuación describimos algunos métodos existentes de selección de variables para este problema.

Jolliffe

Jolliffe (2002) propone diversos métodos para seleccionar un subconjunto pequeño de variables d ($d < p$) ad hoc al análisis de componentes principales. Entre ellos podemos destacar:

J1. Asocia una variable con cada una de las últimas $p-d$ componentes principales. Luego, elimina la variable con mayor peso en valor absoluto que todavía no haya sido seleccionada empezando por la última componente principal y retiene las restantes. Este procedimiento puede ser aplicado en un solo paso o iterativamente. En el segundo criterio, en el primer paso se eliminan e_1 variables ($e_1 < p - d$) y luego se re–calculan las

componentes con las $p - e_1$ variables restantes. En una segunda etapa se eliminan e_2 variables ($e_1 + e_2 \leq p - d$) y así sucesivamente hasta eliminar $p - d$ variables.

J2. Asocia una variable a cada una de las d primeras componentes principales, la variable con mayor peso en valor absoluto que todavía no haya sido seleccionada, y retiene esas variables. Al igual que en el método J1, esto puede ser realizado en un solo paso o iterativamente.

McCabe

McCabe (1984) presenta una estrategia diferente que consiste en seleccionar un conjunto de las variables originales sin pasar por el análisis de componentes principales.

Sea $\mathbf{X} \in \mathbb{R}^p$ un vector aleatorio $I \subset \{1, \ldots, p\}$ un conjunto de índices con cardinal d, sea $\mathbf{X}(I) \in \mathbb{R}^d$ el vector aleatorio formado por las variables del vector $\mathbf{X}$ indicadas por el conjunto I y $\mathbf{X}(I^c) \in \mathbb{R}^{p-d}$ el vector formado por las variables que pertenecen al vector $\mathbf{X}$ y no a $\mathbf{X}(I)$. Denotamos $\Sigma_{\mathbf{X}}$ la matriz de covarianza del vector $\mathbf{X}$, $\Sigma_{\mathbf{X}(I)}$ la matriz de covarianza del vector $\mathbf{X}(I)$ y $\Sigma_{\mathbf{X}(I^c)|\mathbf{X}(I)}$ la matriz de covarianza condicional del vector $\mathbf{X}(I^c)$ dado $\mathbf{X}(I)$.

McCabe (1984) propone elegir las variables indicadas por el conjunto I que resuelvan el siguiente problema de optimización

$$\text{minimizar} \prod_{j=1}^{p-d} \lambda_j, \tag{3.1}$$

donde λ_j con $j = 1, \ldots, p - d$ son los autovalores de la matriz de covarianza condicional $\Sigma_{\mathbf{X}(I^c)|\mathbf{X}(I)}$.

Notemos $\det(A)$ al determinante de la matriz A. Como $\det(\Sigma_{\mathbf{X}(I^c)|\mathbf{X}(I)}) = \prod_{j=1}^{p-d} \lambda_j$, el criterio (3.1) es equivalente a

$$\text{minimizar } \det\left(\Sigma_{\mathbf{X}(I^c)|\mathbf{X}(I)}\right).$$

Observemos que,

$$\det(\Sigma_{\mathbf{X}}) = \det\left(\Sigma_{\mathbf{X}(I)}\right) \det\left(\Sigma_{\mathbf{X}(I^c)|\mathbf{X}(I)}\right),$$

y para cada vector aleatorio $\mathbf{X}$ la matriz $\Sigma_{\mathbf{X}}$ está fija, luego $\det(\Sigma_{\mathbf{X}})$ es un número fijo, implicando que el criterio (3.1) sea equivalente a

$$\text{maximizar } \det\left(\Sigma_{\mathbf{X}(I)}\right).$$

Este nuevo enfoque hace posible que sea computacionalmente factible explorar todos los subconjuntos I de cardinal d.

SCoTLASS

Jolliffe et al. (2003) proponen "Simplified Component Technique - LASSO" (SCoTLASS), cuyo objetivo es seleccionar variables en componentes principales extendiendo la técnica para regresión lineal LASSO.

Su propuesta consiste en calcular los pesos de las componentes principales como en (2.3) con la siguiente restricción adicional,

$$\|\alpha\|_1 = \sum_{i=1}^{p} |\alpha[i]| \leq t,$$

donde $\alpha[i]$ es la i-ésima coordenada del vector α.

Al agregar esta condición en muchos casos se obtienen pesos esparzos para las componentes principales en forma análoga a lo que sucede en el caso de regresión lineal.

SPCA

Zou et al. (2006) proponen "Sparce Principal Component Analysis" (SPCA), para obtener componentes principales con pesos esparzos. Este procedimiento busca extender las ideas de "Elastic Net" para el modelo lineal.

Primero introducen un método para seleccionar variables en dos pasos. Sean $\mathbf{X}_1, \ldots, \mathbf{X}_n \in \mathbb{R}^p$ una muestra aleatoria, en el primer paso calculan los pesos de las componentes principales como en (2.3), en un segundo paso para cada α_k fijo buscan

$$\beta_k = arg\min_{\beta} \sum_{j=1}^{n} \left(\alpha_k' \mathbf{X}_j - \beta' \mathbf{X}_j\right)^2 \text{ sujeto a } \|\beta\|_1 = \sum_{i=1}^{d} |\beta[i]| \leq t_1 \text{ y } \|\beta\|_2^2 = \sum_{i=1}^{d} \beta[i]^2 \leq t_2,$$

donde t_1 y t_2 son constantes, que es equivalente a buscar,

$$\beta_k = arg\min_{\beta} \sum_{j=1}^{n} \left(\alpha_k' \mathbf{X}_j - \beta' \mathbf{X}_j\right)^2 + \lambda_1 \sum_{i=1}^{d} |\beta[i]| + \lambda_2 \sum_{i=1}^{d} \beta[i]^2,$$

con λ_1 y λ_2 positivos. Finalmente definen los pesos de la k-ésima componente principal esparza como el vector β_k normalizado.

Podemos observar que el segundo paso sigue las ideas del método "Naïve Elastic Net" propuesto por Zou & Hastie (2005), el cual difiere del método "Elastic Net" en que a la solución se la multiplica por $(1 + \lambda_2)$. Pero como para hallar las componentes principales esparzas se considera el vector normalizado, ambos métodos obtienen la misma solución.

En una segunda propuesta calculan las componentes principales y las componentes esparzas en simultáneo, a este criterio le dan el nombre de SPCA. Para hallar las primeras l

componentes principales, proponen buscar matrices $\mathbf{A}_{p\times l} \in \mathbb{R}^{p\times l}$ y $\mathbf{B}_{p\times l} \in \mathbb{R}^{p\times l}$ que resuelvan el siguiente problema de optimización,

$$\begin{aligned} &\text{minimizar } \sum_{j=1}^{n} \left\|\mathbf{X}_j - \mathbf{A}\mathbf{B}'\mathbf{X}_j\right\|^2 + \lambda_2 \sum_{i=1}^{l} \|\beta_i\|^2 + \sum_{i=1}^{l} \lambda_{1,i} \|\beta_i\|_1 \\ &\text{sujeto a } \mathbf{A}'\mathbf{A} = I_{l\times l}, \end{aligned}$$

donde $\mathbf{A}_{p\times l}$ es la matriz cuyas columnas son los pesos de las l primeras componentes principales y $\mathbf{B}_{p\times l}$ la matriz cuyas columnas son los pesos de las l primeras componentes principales esparzas e $I_{l\times l} \in \mathbb{R}^{l\times l}$ es la matriz identidad. Se puede observar que el mismo λ_2 es utilizado para las l componentes, pero que se permiten diferentes penalizaciones, $\lambda_{1,i}$ para cada componente esparza.

SPC

Witten et al. (2009) presentan SPC, un método para obtener componentes principales esparzas.

Sea $\mathbf{M}$ la matriz cuya j-ésima fila es el vector aleatorio $\mathbf{X}_j$, y sin pérdida de generalidad supondremos sus columnas centradas. Para hallar los pesos de la primera componente principal esparza proponen buscar $\alpha_1 \in \mathbb{R}^d$ que resuelva el siguiente problema de optimización,

$$\begin{aligned} &\text{maximizar } \alpha'\mathbf{M}'\mathbf{M}\alpha \\ &\text{sujeto a } \|\alpha\|_2^2 \leq 1, \|\alpha\|_1 \leq t. \end{aligned}$$

Notemos que son las mismas condiciones pedidas por SCoTLASS. Lo innovador es como hallar la solución del problema. Proponen aproximar a la matriz $\mathbf{M}$ por una matriz $\widehat{\mathbf{M}}$ de la forma $\widehat{\mathbf{M}} = c\mathbf{u}\mathbf{v}'$, que minimice $\left\|\mathbf{M} - \widehat{\mathbf{M}}\right\|_F^2$ ($\|\cdot\|_F^2$ es la norma de Forbenius al cuadrado, es decir, la suma cuadrática de todos los elementos de la matriz) sujeto a un conjunto de penalidades de $\mathbf{u}$ y $\mathbf{v}$ y c es una constante no negativa. Este modo de descomponer a la matriz M se llama "Penalized Matrix Descomposition" (PMD).

Para calcular las siguientes componentes utilizan un criterio diferente al propuesto por SCoTLASS. Definen la matriz $\mathbf{M}^{k+1} = \mathbf{M}^k - c_k\mathbf{u}_k\mathbf{v}_k'$ donde $\mathbf{M}^1 = \mathbf{M}$ y $\widehat{\mathbf{M}^k} = c_k\mathbf{u}_k\mathbf{v}_k'$. Luego para calcular los pesos de la $k+1$-ésima componente principal esparza buscan el vector que resuelva el siguiente problema de optimización

$$\begin{aligned} &\text{maximizar } \alpha'\mathbf{M'}^{k+1}\mathbf{M}^{k+1}\alpha \\ &\text{sujeto a } \|\alpha\|_2^2 \leq 1, \|\alpha\|_1 \leq t. \end{aligned}$$

Una gran ventaja de esta nueva forma de resolver el problema es que es muy eficiente computacionalmente.

3.2 Procedimiento *Blinding* Componentes Principales

3.2.1 Versión Poblacional

A continuación damos una nueva propuesta que llamaremos Procedimiento *Blinding* Componentes Principales (CP1), para encontrar un subconjunto de las variables originales que explique la salida de componentes principales del mejor modo posible. Recordemos que en (2.3) definimos los pesos de la k-ésima componente principal α_k. Para indicar que estos pesos dependen de la distribución del vector original $\mathbf{X}$ los denotamos $\alpha_k(P)$, donde P es la distribución del vector $\mathbf{X}$. Definimos para cada $I \in \mathcal{I}_d$, $\alpha_k(Q(I))$ el vector de los pesos de la k-ésima componente principal del vector *blinded* $\mathbf{Z}(I)$ definido por la ecuación (2.7).

Supongamos que las primeras $l < p$ componentes principales son suficientes para tener una buena representación de los datos originales. El objetivo es encontrar un conjunto I tal que $\alpha_k(Q(I))$ esté lo más cerca posible de $\alpha_k(P)$ para todo $k = 1, \ldots, l$.

La noción de cercanía la damos con la siguiente función objetivo,

$$h(I) = \sum_{k=1}^{l} p_k \left\| \alpha_k(P) - \alpha_k(Q(I)) \right\|^2, \tag{3.2}$$

con $p_k \geq 0$ y $\sum_{k=1}^{l} p_k = 1$, donde p_k es el peso que le asignamos a la distancia entre las k-ésimas componentes principales. Si consideramos que las primeras l componentes principales son igualmente importantes, recomendamos tomar $p_k = 1/l$. Otra opción es considerar a los pesos proporcionales a la varianza que explica cada componente, es decir $p_k = \lambda_k / \sum_{k=1}^{l} \lambda_k$. Observemos que la función objetivo (3.2) es un promedio ponderado de las distancias al cuadrado entre los pesos de las componentes principales del vector original y del vector *blinded*.

Dado $d < p$ buscamos un conjunto $I \in \mathcal{I}_d$ que minimice la función objetivo (3.2), es decir

$$\mathcal{I}_0 = arg \min_{I \in \mathcal{I}_d} h(I). \tag{3.3}$$

Observación 3 *En Fraiman et al. (2008) la función objetivo mide la similitud entre los resultados de procesos estadísticos, mientras que en este caso mide la discrepancia.*

Observación 4 *Como mencionamos anteriormente, en presencia de outliers es más adecuado considerar un procedimiento robusto en la definición del vector blinded. En (2.7) se puede considerar la mediana condicional, med*$(X[i]|\boldsymbol{X}(I))$ *en lugar de la esperanza condicional,* $E(X[i]|\boldsymbol{X}(I))$.

Si en lugar de buscar un conjunto de variables que expliquen las primeras l componentes principales se desea saber cuales son las variables que mejor representan a cierta componente principal se puede definir la función objetivo como,

$$h^k(I) = \left\| \alpha_k(P) - \alpha_k(Q(I)) \right\|^2, \tag{3.4}$$

y buscar un conjunto $I \in \mathcal{I}_d$ que minimice (3.4).

3.2.2 Versión Empírica

Para describir la versión empírica necesitamos estimar en forma consistente al conjunto I_0, $I_0 \in \mathcal{I}_0$, basado en la muestra de vectores aleatorios $\mathbf{X}_1, \ldots, \mathbf{X}_n$ con distribución P_n.

Dado un conjunto $I \in \mathcal{I}_d$ el primer paso es obtener la versión *blinded* de la muestra de vectores aleatorios en $\mathbb{R}^p$, $\hat{\mathbf{X}}_1(I), \ldots, \hat{\mathbf{X}}_n(I)$, que solo dependan de las coordenadas indicadas por el conjunto I, estimando la esperanza condicional con un estimador no paramétrico.

A modo de ejemplo, consideramos el estimador de vecinos más cercanos (r-NN). Fijamos un entero r, la cantidad de vecinos, y calculamos entre las observaciones $\mathbf{X}_1, \ldots, \mathbf{X}_n$ la distancia Euclídea restringida a las coordenadas de I, es decir las distancias entre los vectores $\mathbf{X}_j(I)$, con $j = 1, \ldots, n$. Para cada $j \in \{1, \ldots, n\}$, encontramos el conjunto de índices C_j de los r vecinos más cercanos de $\mathbf{X}_j(I)$ entre $\{\mathbf{X}_1(I), \ldots, \mathbf{X}_n(I)\}$.

Luego definimos los vectores aleatorios $\hat{\mathbf{X}}_j(I)$ para $j = 1, \ldots, n$ del siguiente modo,

$$\begin{aligned} \hat{X}_j(I)[i] &= \begin{cases} X_j[i] & \text{si } i \in I, \\ \frac{1}{r}\sum_{m \in C_j} X_m[i] & \text{si } i \notin I, \end{cases} \\ &= \begin{cases} X_j[i] & \text{si } i \in I, \\ \frac{1}{r}\sum_{k=1}^{n} X_k[i] I_{\{\|\mathbf{X}_k(I) - \mathbf{X}_j(I)\| \leq R_j(I)\}} & \text{si } i \notin I, \end{cases} \end{aligned} \tag{3.5}$$

donde $X_j[i]$ es la i-ésima coordenada del vector $\mathbf{X}_j$ y $R_j(I)$ es la distancia de $\mathbf{X}_j(I)$ a su r-ésimo vecino más cercano.

A la distribución de $\left\{\hat{\mathbf{X}}_j(I), 1 \leq j \leq n\right\}$ la denotamos $Q_n(I)$.

Observación 5 *En lugar de distancia Euclídea, se puede considerar la distancia de Mahalanobis que es invariante a la escala de los datos.*

Dado un conjunto de índices $I \in \mathcal{I}_d$, definimos la versión empírica de la función objetivo (3.2), como

$$h_n(I) = \sum_{k=1}^{l} p_k \left\| \alpha_k^n(P_n) - \alpha_k^n(Q_n(I)) \right\|^2, \tag{3.6}$$

donde $\alpha_k^n(P_n)$ y $\alpha_k^n(Q_n(I))$ son los pesos de la k-ésima componente principal de los vectores aleatorios $\{\mathbf{X}_1, \ldots, \mathbf{X}_n\}$ y $\left\{\hat{\mathbf{X}}_1(I), \ldots, \hat{\mathbf{X}}_n(I)\right\}$ respectivamente.

Luego, elegimos los conjuntos de variables que minimicen la función (3.6), es decir,

$$\mathcal{I}_n = arg \min_{I \in \mathcal{I}_d} h_n(I). \tag{3.7}$$

A continuación enunciamos las hipótesis necesarias para probar la consistencia del criterio propuesto.

H1. Para todo $i \notin I$, sea $\eta_n^i(z)$ un estimador no paramétrico fuertemente consistente de $\eta^i(z) = E(X[i]|\mathbf{X}(I) = z)$ para casi todo z (P), es decir, $\eta_n^i(z) \to_{c.s.} \eta^i(z)$. Las condiciones bajo las cuales se cumple **H1** se encuentran en Devroye (1981, 1982).

HP1. $E\left(\|\mathbf{X}\|^2\right) < \infty$. La matriz de covarianza del vector aleatorio $\mathbf{X}$ es definida positiva y sus autovalores son todos diferentes.

Teorema 1 *Sean* $\left\{\mathbf{X}_j, j \geq 1\right\}$ *vectores aleatorios independientes e idénticamente distribuidos de dimensión p que satisfacen (2.4). Dado d,* $1 \leq d \leq p$*, sea* $\mathcal{I}_d$ *la familia de todos los subconjuntos de* $\{1, \ldots, p\}$ *con cardinal d y sea* $\mathcal{I}_0 \subset \mathcal{I}_d$ *la familia de todos los subconjuntos para los cuales se alcanza el mínimo de la ecuación (3.2). Bajo* **H1** *y* **HP1**, *tenemos que dado* $I_n \in \mathcal{I}_n$*, existe un* $n_0(\omega)$ *tal que, para todo* $n \geq n_0(\omega)$*, con probabilidad uno,* $I_n \in \mathcal{I}_0$.

La demostración del Teorema 1 se encuentra en el último capítulo.

3.3 Consideraciones Prácticas

En esta sección damos algunas consideraciones para implementar el método propuesto. En primer lugar describimos un procedimiento para obtener el número de vecinos en forma consistente al calcular la esperanza condicional. En segundo lugar, damos una regla para elegir la cantidad de variables (d).

3.3.1 Una estimación no paramétrica para la esperanza condicional

Dado un conjunto $I \in \mathcal{I}_d$ en (3.5) definimos los vectores aleatorios $\hat{\mathbf{X}}_j(I)$ para una cantidad r fija de vecinos más cercanos. A continuación explicamos una forma de elegir de modo consistente a r. Primero notemos que para cada conjunto I y para cada coordenada $i \notin I$ hay que elegir la cantidad de vecinos más cercanos, es decir, $r = r(i, I)$. Para estimar este número sugerimos utilizar el método de validación cruzada propuesto por Li & Gong (1987), en el cual estiman en forma consistente el número de vecinos. Ellos proponen elegir

$$\widehat{r}_{opt}(i, I) = arg \min_r \frac{\frac{1}{n}\sum_{k=1}^n \left(X_k[i] - \widehat{X}_k[i]\right)^2}{\left(1 - \frac{1}{n}\,\text{traza}(M_n(r))\right)^2},$$

donde $(M_n(r))_{ij} = W_{ij}/\sum_{l=1}^n W_{il}$ siendo $W_{ij} = W\left(\left(\mathbf{X}_k(I) - \mathbf{X}_j(I)\right)/R_k(I)\right)$ y $R_k(I)$ es la distancia de $\mathbf{X}_k(I)$ a su r-ésimo vecino más cercano.

Si W es la función uniforme, que le asigna el mismo peso a cada uno de los r vecinos más cercanos de $X_k(I)$, tenemos,

$$\begin{aligned} W_{ij} &= \begin{cases} 1 & \text{si } \left\| \frac{\mathbf{X}_k(I)-\mathbf{X}_j(I)}{R_k(I)} \right\| \leq 1, \\ 0 & \text{caso contrario,} \end{cases} \\ &= \begin{cases} 1 & \text{si } \|\mathbf{X}_k(I) - \mathbf{X}_j(I)\| \leq R_k(I), \\ 0 & \text{caso contrario.} \end{cases} \end{aligned}$$

En este caso $(M_n(r))_{ii} = W_{ii} / \sum_{l=1}^{n} W_{il} = 1/r$, implicando que traza$(M_n(r)) = n/r$, luego

$$\widehat{r}_{opt}(i, I) = arg \min_r \frac{\frac{1}{n} \sum_{k=1}^{n} \left(X_k[i] - \widehat{X}_k[i] \right)^2}{\left(1 - \frac{1}{r}\right)^2}.$$

Es sabido que r-NN no puede ser aplicado de forma eficaz cuando hay variables categóricas. Aquí hay que puntualizar dos situaciones. La variable categórica, puede pertenecer al conjunto de las variables explicativas (las designadas por el conjunto I) o pertenecer al conjunto de las variables que hay que predecir (no pertenecer al conjunto I). En el primer caso, proponemos buscar los r vecinos más cercanos para el dato $\mathbf{X}_j$, entre las observaciones para las cuales la variable categórica (o las variables categóricas, si hay más de una) tome el mismo valor. En el segundo caso, es decir cuando hay que estimar esta variable, proponemos asignarle como valor la moda de los r vecinos más cercanos.

3.3.2 Un método para decidir cuantas variables seleccionar

Otra cuestión importante es como elegir la cantidad de variables con las cuales quedarse, es decir d. Por un lado, la función objetivo h decrece cuando d crece y por el otro, buscamos un valor chico de h para un conjunto I de cardinal pequeño. Recordemos que h es un promedio ponderado de las distancias al cuadrado entre los pesos de las primeras l componentes principales de la variable original y los de la variable *blinded*. Como los pesos de las componentes tienen norma uno, existe una relación directa entre la distancia al cuadrado que los separa ($\|\alpha_k(P) - \alpha_k(Q(I))\|^2$) y el ángulo comprendido entre los dos vectores ($\varphi(\alpha_k(P), \alpha_k(Q(I)))$). Se ve claramente que cuanto más chico sea el ángulo $\varphi(\alpha_k(P), \alpha_k(Q(I)))$ menor es $\|\alpha_k(P) - \alpha_k(Q(I))\|^2$. Luego, proponemos fijar un ángulo γ y elegir d el entero más chico que cumpla

$$\varphi(\alpha_k(P), \alpha_k(Q(I))) \leq \gamma \text{ para todo } k = 1, \ldots, l.$$

3.4 Ejemplo Real: Vertebral Column Data Set

A continuación ilustramos el comportamiento del método propuesto analizando el conjunto de datos "Vertebral Column Data Set" del repositorio de la Universidad de California, Irvine

(Frank & Asuncion (2010)). Este conjunto contiene seis características biomecánicas que fueron usadas para clasificar a un grupo de pacientes en tres grupos, aquellos que tienen hernia de disco, otro con espondilolistesis y por último pacientes sanos. Cada categoría cuenta con 60, 50 y 100 observaciones respectivamente. A cada uno de los pacientes se les midieron 6 características que describen la orientación y la forma de la pelvis y la columna lumbar. Las características medidas en cada uno de los pacientes son (a) incidencia pélvica, (b) inclinación pélvica, (c) ángulo de lordosis lumbar, (d) pendiente sacra, (e) radio pélvico y (f) grado de espondilolistesis.

Las dos primeras componentes explican el 85% de la varianza, luego aplicamos el procedimiento propuesto (Procedimiento *Blinding* CP1) asignándole el mismo peso a las dos componentes en la ecuación (3.2), es decir $p_k = 1/2$ con $k = 1, 2$. Al elegir una variable selecciona grado de espondilolistesis y el ángulo más grande que se obtiene es de 7.5 grados, luego decidimos quedarnos solo con dicha variable. El número de vecinos más cercanos fue cross–validado y resultó ser 55 para las variables (a) y (b), 70 para la variable (c), 102 para la variable (d) y 39 para la variable (e), obteniendo $h_n(I) = 0.017$. Realizamos el procedimiento considerando la distancia Euclídea, pero observemos que para $d = 1$ esta coincide con la distancia de Mahalanobis.

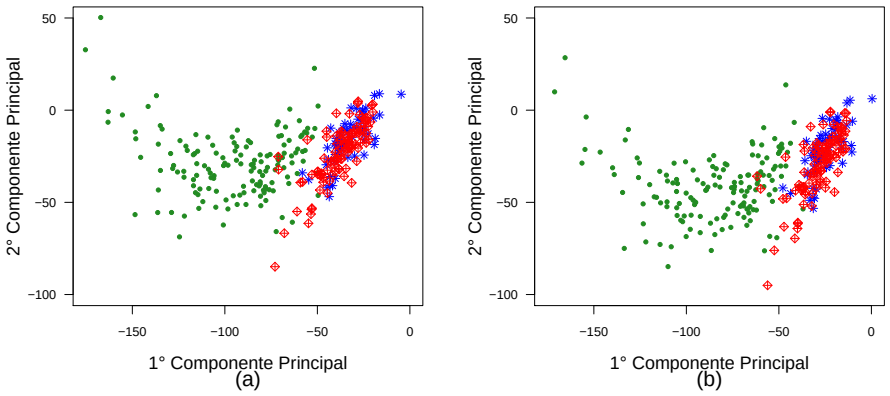

Figura 3.1: (a) Proyecciones de los datos en el vector de los pesos de las dos primeras componentes principales originales. (b) Proyecciones de los datos en el vector de los pesos de las dos primeras componentes principales del Procedimiento *Blinding* CP1. Estrella azul: Paciente con hernia de disco. Punto verde: Paciente con espondilolistesis. Diamante rojo: Paciente sano.

El gráfico que obtenemos al proyectar los datos en las dos primeras componentes principales es muy similar al gráfico cuando se proyectan los datos en las componentes principales calculadas usando el Procedimiento *Blinding* CP1 (Figura 3.1). En los gráficos también observamos que los pacientes con espondilolistesis están separados del resto de los pacientes,

pero los pacientes con hernia de disco y los pacientes sanos están entremezclados.

Con el objetivo de comprender cuales son las variables a través de las cuales se diferencian estas dos categorías de pacientes. Realizamos una segunda etapa. Consideramos los pacientes con hernia de disco y los sanos, para ellos realizamos el análisis con las dos primeras componentes principales porque explican el 75% de la varianza. Al aplicar el Procedimiento *Blinding* CP1 para elegir una sola variable el ángulo más grande que se obtiene es de 78 grados, al elegir dos variables de 21 grados y para tres de 19 grados. Claramente hay una mejoría importante cuando se consideran 2 variables en lugar de una, pero este patrón es mucho menos marcado si se retienen 3 variables. Teniendo en cuenta que siempre existe un compromiso entre el número de variables seleccionadas y la bondad de ajuste del modelo, creemos conveniente optar por un modo más parsimonioso reteniendo 2 variables que son ángulo de lordosis lumbar y radio pélvico. Nuevamente no hay diferencias al considerar la distancia Euclídea ($h_n(I) = 0.125$) o la distancia de Mahalanobis ($h_n(I) = 0.141$). En la Figura 3.2 observamos que los gráficos obtenidos al calcular las componentes principales mediante el Procedimiento *Blinding* CP1 considerando ambas distancias son muy similares a los que se obtienen con las componentes calculadas por el procedimiento tradicional.

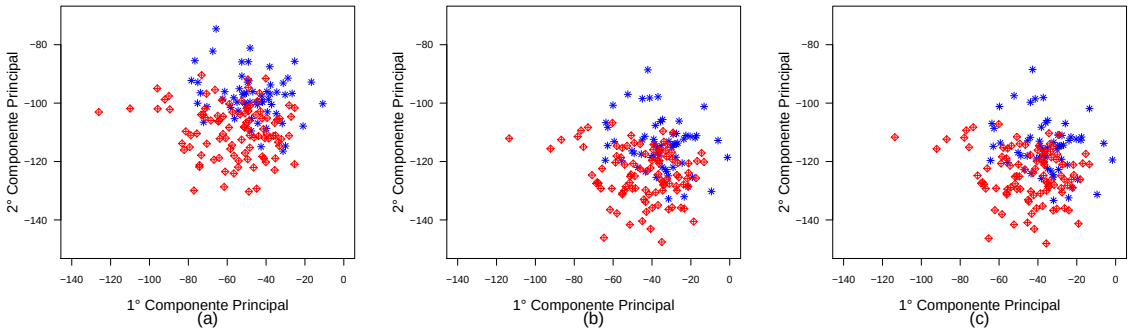

Figura 3.2: (a) Proyecciones de los datos en el vector de los pesos de las dos primeras componentes principales originales. (b) Proyecciones de los datos en el vector de los pesos de las dos primeras componentes principales del Procedimiento *Blinding* CP1 usando la distancia Euclídea. (c) Proyecciones de los datos en el vector de los pesos de las dos primeras componentes principales del Procedimiento *Blinding* CP1 usando la distancia Mahalanobis. Estrella azul: Paciente con hernia de disco. Diamante rojo: Paciente sano.

Capítulo 4

El Problema de Selección de Variables

En este capítulo hacemos una propuesta general para seleccionar variables en concordancia con el criterio introducido por Fraiman et al. (2008). Mostramos la consistencia de los procedimientos propuestos e ilustramos su desempeño mediante simulaciones y análisis de conjuntos de datos reales.

4.1 Procedimiento *Blinding* Multivariado

En esta sección presentamos el Procedimiento *Blinding* Multivariado (M) que es una propuesta general para seleccionar variables. Siguiendo las mismas ideas utilizadas para describir el Procedimiento *Blinding* CP1 definimos para diferentes modelos estadísticos una función objetivo $h(I)$. Esta función mide la bondad de predicción de las variables indicadas por el conjunto I. Del mismo modo que en (3.3) seleccionamos un conjunto que minimice la función objetivo. Para el caso de componentes principales damos una nueva propuesta cuyo espíritu se asemeja más a las propuestas para los otros modelos estadísticos.

A partir de la muestra de vectores aleatorios $\mathbf{X}_1, \ldots, \mathbf{X}_n$ y considerando los vectores aleatorios $\hat{\mathbf{X}}_1(I), \ldots, \hat{\mathbf{X}}_n(I)$ definidos en (3.5) damos en cada caso la versión empírica de la función objetivo.

4.1.1 El Modelo de Regresión

Recordemos el modelo de regresión que fue definido en (2.1), si consideramos $\psi(\mathbf{X}, g) = g(\mathbf{X}, \beta_0)$, nuestro objetivo es encontrar un conjunto I tal que $\psi(\mathbf{Z}(I), g) = g(\mathbf{Z}(I), \beta_0)$ esté lo más cerca posible de $\psi(\mathbf{X}, g) = g(\mathbf{X}, \beta_0)$.

En este caso, la función objetivo es:

$$\begin{aligned} h(I) &= E\left((\psi(\mathbf{X}, g) - \psi(\mathbf{Z}(I), g))^2\right) \\ &= E\left((g(\mathbf{X}, \beta_0) - g(\mathbf{Z}(I), \beta_0))^2\right). \end{aligned} \tag{4.1}$$

Observemos que la función (4.1) mide la esperanza de la distancia al cuadrado entre la función de regresión con las variables originales y con el vector *blinded*. Luego, dado $d < p$, buscamos un conjunto $I \in \mathcal{I}_d$ que minimice la función objetivo (4.1), como en (3.3).

La versión empírica de la función (4.1) está dada por

$$h_n(I) = \frac{1}{n}\sum_{j=1}^{n}\left(g\left(\mathbf{X}_j, \beta_n\right) - g\left(\hat{\mathbf{X}}_j(I), \beta_n\right)\right)^2. \tag{4.2}$$

Elegimos los conjuntos de variables que minimicen la función (4.2), del mismo modo que en (3.7).

Para establecer el resultado de consistencia además de utilizar la condición **H1** definida en el capítulo anterior necesitamos los siguientes supuestos:

HR1. Sea β_n un estimador fuertemente consistente de β_0 ($\|\beta_n - \beta_0\| \to_{c.s.} 0$) y g una función de regresión continua.

HR2. $E\left(g^2\left(\mathbf{X}, \beta_0\right)\right) < \infty$.

Teorema 2 *Sean* $\left\{\left(\mathbf{X}_j, Y_j\right), j \geq 1\right\}$ *vectores aleatorios independientes e idénticamente distribuidos de dimensión* $p+1$ *que satisfacen (2.1). Dado* $d, 1 \leq d \leq p$, *sea* $\mathcal{I}_d$ *la familia de todos los subconjuntos de* $\{1, \ldots, p\}$ *con cardinal* d *y sea* $\mathcal{I}_0 \subset \mathcal{I}_d$ *la familia de todos los subconjuntos para los cuales se alcanza el mínimo de la ecuación (4.1). Bajo* **H1**, **HR1** *y* **HR2**, *tenemos que para cada* $I_n \in \mathcal{I}_n$, *existe un* $n_0(\omega)$ *tal que, para todo* $n \geq n_0(\omega)$, *con* $I_n \in \mathcal{I}_0$, *con probabilidad uno.*

La demostración del Teorema 2 se encuentra en el último capítulo.

4.1.2 Modelo Lineal Generalizado

Recordemos la definición del modelo lineal generalizado que fue dada en (2.2). Si consideramos $\psi(\mathbf{X}, g) = g^{-1}(\mathbf{X}'\beta_0)$, nuestro objetivo es encontrar un conjunto I tal que $\psi(\mathbf{Z}(I), g) = g^{-1}(\mathbf{Z}(I)'\beta_0)$ esté lo más cerca posible de $\psi(\mathbf{X}, g) = g^{-1}(\mathbf{X}'\beta_0)$.

En este caso, la función objetivo es,

$$\begin{aligned} h(I) &= E\left(\left(\psi(\mathbf{X}, g) - \psi(\mathbf{Z}(I), g)\right)^2\right) \\ &= E\left(\left(g^{-1}\left(\mathbf{X}'\beta_0\right) - g^{-1}\left(\mathbf{Z}(I)'\beta_0\right)\right)^2\right). \end{aligned} \tag{4.3}$$

Esta función mide la esperanza de la distancia al cuadrado entre la inversa de la función de link con las variables originales y con las variables *blinded*. Luego, dado $d < p$, buscamos un conjunto $I \in \mathcal{I}_d$ que minimice la función objetivo (4.3).

La versión empírica de la función (4.3) está dada por

$$h_n(I) = \frac{1}{n}\sum_{j=1}^{n}\left(g^{-1}\left(\mathbf{X}_j'\beta_n\right) - g^{-1}\left(\hat{\mathbf{X}}_j(I)'\beta_n\right)\right)^2. \tag{4.4}$$

Elegimos los conjuntos de variables que minimicen la función (4.4), del mismo modo que en (3.7).

Para establecer el resultado de consistencia, además de necesitar la hipótesis **H1** ya definida, necesitamos las siguientes condiciones adicionales.

HG1. Sea β_n un estimador fuertemente consistente de β_0 ($\|\beta_n - \beta_0\| \to_{c.s.} 0$) y g^{-1} continua.

HG2. $E\left(\left(g^{-1}\left(\mathbf{X}, \beta_0\right)\right)^2\right) < \infty$.

Teorema 3 *Sean* $\left\{\left(\mathbf{X}_j, Y_j\right), j \geq 1\right\}$ *vectores aleatorios independientes e idénticamente distribuidos de dimensión* $p+1$ *que satisfacen (2.2). Dado* $d, 1 \leq d \leq p$, *sea* $\mathcal{I}_d$ *la familia de todos los subconjuntos de* $\{1, \ldots, p\}$ *con cardinal* d *y sea* $\mathcal{I}_0 \subset \mathcal{I}_d$ *la familia de todos los subconjuntos para los cuales se alcanza el mínimo de la ecuación (4.3). Bajo* **H1**, **HG1** *y* **HG2**, *tenemos que para cada* $I_n \in \mathcal{I}_n$, *existe un* $n_0(\omega)$ *tal que, para todo* $n \geq n_0(\omega)$, *con* $I_n \in \mathcal{I}_0$, *con probabilidad uno.*

La demostración del Teorema 3 es análoga a la del Teorema 2 intercambiando simplemente g por g^{-1}.

4.1.3 Componentes Principales

Recordemos que en (2.3) definimos α_k el vector de los pesos de la k-ésima componente principal y en (2.4) denotamos $U_k = \alpha_k'\mathbf{X}$ a la componente principal. Definimos para cada $I \in \mathcal{I}_d$, $U_k(I) = \alpha_k'\mathbf{Z}(I)$.

Asumimos que las primeras $l < p$, componentes principales son suficientes para tener una buena representación de los datos originales. Luego, nuestro objetivo es encontrar un conjunto I tal que $U_k(I)$ esté lo más cerca posible de U_k para todo $k = 1, \ldots, l$.

Definimos a la función objetivo del siguiente modo,

$$h(I) = \sum_{k=1}^{l} E\left((U_k - U_k(I))^2\right). \tag{4.5}$$

Esta función mide la suma de las esperanzas de las distancias al cuadrado entre las proyecciones de las primeras l componentes principales, considerando las variables originales y considerando las variables *blinded*. Dado $d < p$, buscamos un conjunto $I \in \mathcal{I}_d$ que minimice la función objetivo (4.5), como en (3.3).

La versión empírica de la función (4.5) está dada por

$$h_n(I) = \sum_{k=1}^{l} \frac{1}{n} \sum_{j=1}^{n} \left(\alpha_k'^n \mathbf{X}_j - \alpha_k'^n \hat{\mathbf{X}}_j(I) \right)^2 . \tag{4.6}$$

Elegimos los conjuntos de variables que minimicen la función (4.6), del mismo modo que en (3.7).

Para dar el resultado de consistencia, además de necesitar las hipótesis **H1** y **HP1**, pedimos la siguiente condición adicional.

HP2. $E\left((Z(I)[i] - X(I)[i])^2\right) < \infty$ para $1 \leq i \leq p$.

Teorema 4 *Sean* $\left\{\mathbf{X}_j, j \geq 1\right\}$ *vectores aleatorios independientes e idénticamente distribuidos de dimensión p que satisfacen (2.4). Dado* $d, 1 \leq d \leq p$, *sea* $\mathcal{I}_d$ *la familia de todos los subconjuntos de* $\{1, \ldots, p\}$ *con cardinal d y sea* $\mathcal{I}_0 \subset \mathcal{I}_d$ *la familia de todos los subconjuntos para los cuales se alcanza el mínimo de la ecuación (4.5). Bajo* **H1**, **HP1** *y* **HP2**, *tenemos que para cada* $I_n \in \mathcal{I}_n$, *existe un* $n_0(\omega)$ *tal que, para todo* $n \geq n_0(\omega)$, *con* $I_n \in \mathcal{I}_0$, *con probabilidad uno.*

La demostración del Teorema 4 se encuentra en el último capítulo.

4.1.4 Correlación Canónica

Recordemos que en (2.5) definimos α_k y β_k los vectores que nos determinan los pesos de las k-ésimas variables canónicas y en (2.6) a las variables canónicas $U_k = \alpha_k'\mathbf{X}$ y $V_k = \beta_k'\mathbf{Y}$. Definimos para cada $I_{\mathbf{X}} \in \mathcal{I}_{d_{\mathbf{X}}}$ e $I_{\mathbf{Y}} \in \mathcal{I}_{d_{\mathbf{Y}}}$, $U_k(I_{\mathbf{X}}) = \alpha_k'\mathbf{Z}(I_{\mathbf{X}})$ y $V_k(I_{\mathbf{Y}}) = \alpha_k'\mathbf{Z}(I_{\mathbf{Y}})$.

Asumimos que las primeras $l < \widetilde{p}$ variables canónica son suficientes para tener una buena representación de los datos originales. Nuestro objetivo es encontrar un conjunto $I_{\mathbf{X}} \in \mathcal{I}_{d_{\mathbf{X}}}$ y un conjunto $I_{\mathbf{Y}} \in \mathcal{I}_{d_{\mathbf{Y}}}$, tales que $U_k(I_{\mathbf{X}})$ esté lo más cerca posible de U_k y $V_k(I_{\mathbf{Y}})$ esté lo más cerca posible de V_k para todo $k = 1, \ldots, l$.

Buscamos dos conjuntos de variables $I_{\mathbf{X}} \in \mathcal{I}_{d_{\mathbf{X}}}$ e $I_{\mathbf{Y}} \in \mathcal{I}_{d_{\mathbf{Y}}}$ que minimicen nuestras funciones objetivo

$$h\left(I_{\mathbf{X}}\right) = \sum_{k=1}^{l} E\left((U_k - U_k(I_{\mathbf{X}}))^2\right) \text{ y } h(I_{\mathbf{Y}}) = \sum_{k=1}^{l} E\left((V_k - V_k(I_{\mathbf{Y}}))^2\right). \tag{4.7}$$

Al igual que en componentes principales, las funciones objetivo descritas en (4.7) miden la suma de las esperanzas de las distancias al cuadrado entre las proyecciones de las primeras l variables canónicas, considerando las variables originales y considerando las variables *blinded*. Dado $d < \tilde{p}$, buscamos un conjunto $I_{\mathbf{X}} \in \mathcal{I}_{d_{\mathbf{X}}}$ y un conjunto $I_{\mathbf{Y}} \in \mathcal{I}_{d_{\mathbf{Y}}}$ que minimicen las funciones objetivo dadas en (4.7).

La versión empírica de las funciones (4.7) está dada por

$$h_n(I_{\mathbf{X}}) = \sum_{k=1}^{l} \frac{1}{n} \sum_{j=1}^{n} \left(\alpha_k'^n \mathbf{X}_j - \alpha_k'^n \hat{\mathbf{X}}_j(I_{\mathbf{X}})\right)^2 \text{ y } h_n(I_{\mathbf{Y}}) = \sum_{k=1}^{l} \frac{1}{n} \sum_{j=1}^{n} \left(\beta_k'^n \mathbf{Y}_j - \beta_k'^n \hat{\mathbf{Y}}_j(I_{\mathbf{Y}})\right)^2. \quad (4.8)$$

Elegimos los conjuntos de variables que minimicen las funciones dadas en (4.8).

Para demostrar consistencia, necesitamos la condición **H1** y además las siguientes hipótesis adicionales,

HC1. $E\left(\|\mathbf{X}\|^2\right) < \infty$ y $E\left(\|\mathbf{Y}\|^2\right) < \infty$. Las matrices de covarianza $\Sigma_{\mathbf{X}}$ y $\Sigma_{\mathbf{Y}}$ son definidas positivas y todos sus autovalores son diferentes.

HC2. $E\left((Z(I_{\mathbf{X}})[i] - X(I_{\mathbf{X}})[i])^2\right) < \infty$ y $E\left((Z(I_{\mathbf{Y}})[i'] - Y(I_{\mathbf{Y}})[i'])^2\right) < \infty$ para $1 \leq i \leq p$, $1 \leq i' \leq q$.

Teorema 5 *Sean* $\left\{\left(\mathbf{X}_j, \mathbf{Y}_j\right), j \geq 1\right\}$ *vectores aleatorios independientes e idénticamente distribuidos de dimensión* $p+q$ *que satisfacen (2.6). Dado* $d_{\mathbf{X}}, 1 \leq d_{\mathbf{X}} \leq \widetilde{p}$ *y* $d_{\mathbf{Y}}, 1 \leq d_{\mathbf{Y}} \leq \widetilde{p}$*, sea* $\mathcal{I}_{d_{\mathbf{X}}}$ *e* $\mathcal{I}_{d_{\mathbf{Y}}}$ *la familia de todos los subconjuntos de* $\{1, \ldots, p\}$ *con cardinal* $d_{\mathbf{X}}$ *y* $d_{\mathbf{Y}}$ *respectivamente y sean* $\mathcal{I}_{d_{\mathbf{X},0}} \subset \mathcal{I}_{d_{\mathbf{X}}}$ *e* $\mathcal{I}_{d_{\mathbf{Y},0}} \subset \mathcal{I}_{d_{\mathbf{Y}}}$ *las familias de todos los subconjuntos para los cuales se alcanza el mínimo de las ecuaciones (4.7). Bajo* **H1**, **HC1** *y* **HC2**, *tenemos que para cada* $I_{d_{\mathbf{X},n}} \in \mathcal{I}_{d_{\mathbf{X},n}}$ *y para cada* $I_{d_{\mathbf{Y},n}} \in \mathcal{I}_{d_{\mathbf{Y},n}}$*, existe un* $n_0(\omega)$ *tal que, para todo* $n \geq n_0(\omega)$*,* $I_{d_{\mathbf{X},n}} \in \mathcal{I}_{d_{\mathbf{X},0}}$ *e* $I_{d_{\mathbf{Y},n}} \in \mathcal{I}_{d_{\mathbf{Y},0}}$*, con probabilidad uno.*

La demostración del Teorema 5 es análoga a la del Teorema 4.

4.2 Consideración Práctica

En esta sección presentamos un procedimiento heurístico para resolver el problema de como elegir d, la cantidad de variables que consideramos. Para cada $d = \sharp I$ existe un conjunto $I_0(d)$ para el cual la función objetivo h_n se minimiza, y el valor que alcanza es $h_n(I_0(d))$. Es claro que $h_n(I_0(d))$ es una función decreciente en d, es decir, a medida que aumentamos la cantidad de variables $h_n(I_0(d))$ decrece. Nuestra propuesta consiste en mirar el gráfico $(d, h_n(I_0(d)))$ y decidir para que valor de d la poligonal que une los puntos tiene una pendiente pronunciada a la izquierda de d, pero no a la derecha. Es decir, para que valor de d se puede visualizar un codo. Este método de como elegir la cantidad de variables tiene el mismo espíritu que otros métodos que se han usado en problemas de cluster y en análisis de componentes principales (ver Jolliffe (2002)).

Observación 6 *Se podría haber considerado para seleccionar el número de variables un criterio del estilo de AIC o BIC, penalizando a nuestra función objetivo con un término que considere el número de variables seleccionadas.*

4.3 Simulaciones

A continuación realizamos un estudio de simulación para mostrar alcances y limitaciones de los métodos propuestos. Además, comparamos el desempeño con otros procedimientos de selección de variables que se encuentran en la literatura.

4.3.1 Regresión: El clásico Modelo Lineal

Generamos datos con el siguiente modelo lineal,

$$Y_j = 2 + 3X_{1j} - 4X_{2j} + 2X_{3j} - \frac{3}{2}X_{4j} - X_{5j} + \frac{1}{50}X_{6j} + 3X_{7j} + 4X_{8j} - X_{9j} + e_j,\ \ j = 1, \ldots, n,$$

donde X_{1j} y X_{2j} son variables aleatorias independientes normales con esperanza 0 y varianza 4 y las otras variables aleatorias son funciones de estas dos del siguiente modo,

$$X_{ij} = \begin{cases} (X_{1j}X_{2j})^2 & \text{para } i = 3, \\ (X_{1j}X_{2j})^3 Z_{1j} & \text{para } i = 4, \\ X_{1j}^4 Z_{2j} & \text{para } i = 5, \\ (X_{1j}X_{2j})^5 Z_{3j} & \text{para } i = 6, \\ \exp(X_{1j}X_{2j}) & \text{para } i = 7, \\ \sqrt{|X_{1j}X_{2j}|} & \text{para } i = 8, \\ \exp(X_{2j}) Z_{4j} & \text{para } i = 9, \end{cases}$$

donde Z_{ij} con $i = 1, \ldots, 4$ son variables aleatorias independientes normalmente distribuidas centradas en el origen con varianza 0.25, y los errores e_j son variables aleatorias independientes con distribución $N(0, 1)$.

Las variables X_1 y X_2 son las que generan el modelo, las otras variables son transformaciones no lineales con errores de las mismas.

Consideramos muestras de tamaño $n = 200$ y 300 y realizamos 500 réplicas. Por simplicidad, no buscamos el número óptimo de vecinos para cada variable y prefijamos 10 ó 20 vecinos más cercanos para la estimación no paramétrica. Como sugerimos anteriormente elegimos con cuantas variables quedarnos mirando el gráfico de $(d, h_n(I_0(d)))$.

Luego comparamos nuestro método con otros de selección de variables. Consideramos el método de Akaike (AIC), el Criterio de Información de Bayes (BIC) y el R^2 ajustado (ADJR2). También comparamos con tres propuestas más recientes, LASSO (Tibshirani (1996)), utilizando el algoritmo LARS propuesto por Efron et al. (2004), los parámetros de regularización fueron estimados usando validación cruzada de parámetro 10; SCAD (Fan & Li (2001)) y MCP (Zhang (2010)) utilizando el algoritmo propuesto por Brehency & Huang (2011). Reportamos los resultados en la Tabla 4.1.

En la Tabla 4.1 observamos que el promedio del número de variables elegidas por el Procedimiento *Blinding* M es muy cercano al número de variables que generan el modelo. El procedimiento elige dos variables en más de un 94% de las veces. Más aún, las veces

Tabla 4.1: Primera Línea : Promedio de las variables elegidas. Segunda Línea: Mediana del número de variables elegidas. Tercera Línea: % de aciertos = porcentaje de veces que dos variables fueron elegidas. Cuarta Línea: $NVS \leq 5$ = porcentaje de veces que menos de 5 variables fueron elegidas.

Número de Observaciones		Criterio de Selección de Variables							
		Procedimiento *Blinding* M		AIC	BIC	ADJR2	LASSO	SCAD	MCP
		$r = 10$	$r = 20$						
$n = 200$	Media	2.05	2.05	8.39	8.28	8.19	4.64	3.42	4.31
	Mediana	2	2	9	9	9	4	3	4
	% de aciertos	94.50	94.70	0.40	1.00	2.10	14.2	19.7	13.5
	$NVS \leq 5$	100	100	5.30	7.80	10.60	58.4	75.9	52.6
$n = 300$	Media	2.04	2.04	8.25	8.12	7.87	4.03	3.58	4.36
	Mediana	2	2	9	9	9	4	3	4
	% de aciertos	96.30	96.30	0.20	0.40	1.40	14	19	13
	$NVS \leq 5$	100	100	5.10	8.60	14.10	68.5	70.5	49.8

que no elige dos variables elige tres. Por el contrario, los métodos clásicos de selección de variables retienen, en la mayoría de los casos, todas las variables, o eligen descartar una variable. LASSO, SCAD y MCP funcionan mejor ya que retienen menos variables. LARS elige dos variables un 14% de las veces y la mediana de la cantidad de variables es 4, sin embargo en un 4.9% (respec. 6.6%) de las veces han elegido una sola variable para $n = 200$ (respec. $n = 300$). Entre nuestros competidores el método que mejor performa es SCAD, eligiendo el número correcto de variables en más de un 19% de las veces, pero el algoritmo no converge en un 10.4% (respec. 16.5%) de las veces y en un 30.7% (respec. 28.7%) de las veces elige solamente una variable o únicamente el intercept para $n = 200$ (respec. $n = 300$).

La Figura 4.1 muestra los histogramas que nos indican la cantidad de variables que fueron elegidas por los diferentes procedimientos al considerar muestras de tamaño 200. Observamos que el Procedimiento *Blinding* M nunca elige una variable y que no es sensible a la cantidad de vecinos más cercanos considerada. Los métodos LASSO, SCAD y MCP eligen un número de variables cercano al óptimo mientras que los procedimientos clásicos no son capaces de detectar la dependencia no lineal.

4.3.2 Componentes Principales

Comenzamos con un ejemplo sencillo que fue propuesto por Zou et al. (2006). El modelo tiene 10 variables, pero solo depende de 2. Cada una de las 10 variables es una de las dos variables originales más un ruido o una combinación lineal de las variables originales más un ruido. Para ser más específicos, tenemos las variables Y_{1j} que son independientes y están normalmente distribuidas con esperanza cero y varianza 290, y las variables Y_{2j} que son independientes y están normalmente distribuidas con esperanza cero y varianza 300. Se de-

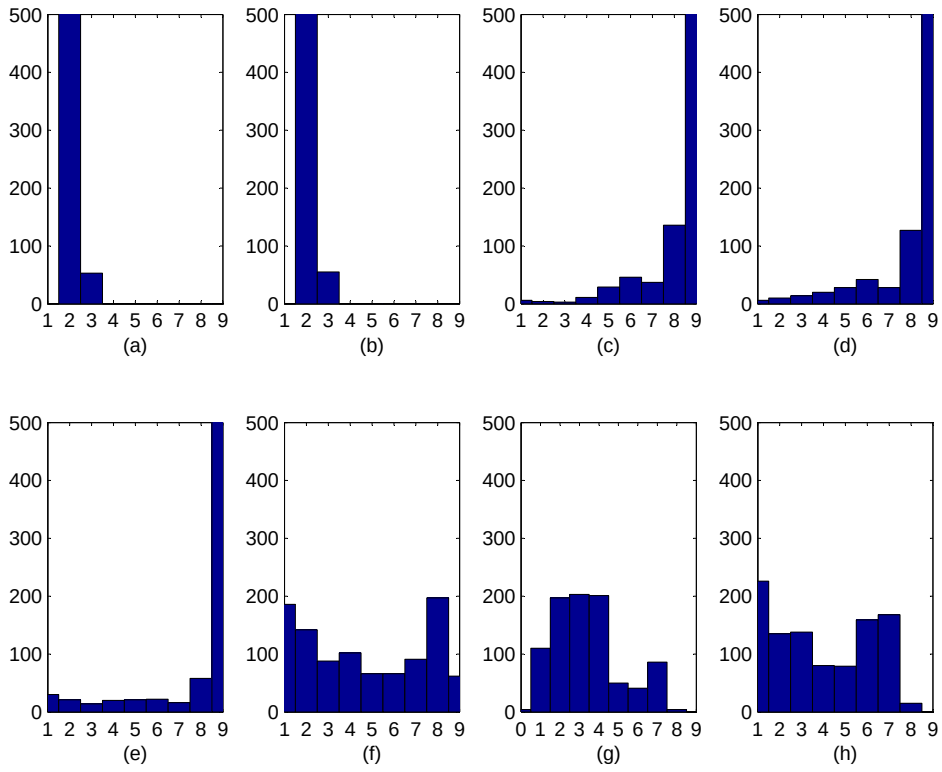

Figura 4.1: Los Histogramas indican el número de veces que cada procedimiento elige d variables con n=200. (a) Procedimiento *Blinding* M con $r = 10$, (b) Procedimiento *Blinding* M con $r = 20$, (c) AIC, (d) BIC, (e) ADJR2, (f) LARS, (g) SCAD, (h) MCP.

finen las variables Y_{3j} como combinaciones lineales de las variables anteriores más un ruido, es decir, $Y_{3j} = -0.3Y_{1j} + 0.925Y_{2j} + e_j$, donde e_j son errores independientes normalmente distribuidos con esperanza cero y varianza unitaria. Las observaciones $\left(X_{1j}, \ldots, X_{10j}\right)$ con $j = 1, \ldots, 100$, están dadas por

$$X_{ij} = \begin{cases} Y_{1j} + e_{ij} & \text{para } i = 1, \ldots, 4, \\ Y_{2j} + e_{ij} & \text{para } i = 5, \ldots, 8, \\ Y_{3j} + e_{ij} & \text{para } i = 9, 10, \end{cases}$$

donde los errores e_{ij} son variables aleatorias $N\left(0, \sigma^2\right)$ con $\sigma^2 = 1$. Las dos primeras componentes principales retienen más de un 99% de la varianza total.

Realizamos 1000 réplicas y aplicamos los dos algoritmos de selección de variables propuestos para el análisis de componentes principales, es decir el Procedimiento *Blinding* CP1 y el Procedimiento *Blinding* M.

Para el Procedimiento *Blinding* CP1 consideramos ángulos de 10, 15, 20 y 25 grados y buscamos por cross–validación la cantidad de vecinos para estimar la esperanza condicional. En la Figura 4.2 se encuentran los resultados para este procedimiento que en un alto porcentaje de las veces elige dos variables. El método funciona mejor para ángulos más chicos, a medida que aumenta el ángulo elige más veces una variable y esto no es bueno. Observemos que el objetivo es retener la información proporcionada por Y_1 e Y_2, por tal motivo un conjunto de variables que sólo dependa de una de ellas es una mala elección. Cuando elige dos variables esto nunca sucede.

Luego aplicamos el Procedimiento *Blinding* M. En este caso además de elegir la cantidad de vecinos para estimar la esperanza condicional por cross–validación, observamos que sucedía al fijar 5, 10 y 20 vecinos más cercanos. En la Figura 4.3 vemos que el valor óptimo de la función objetivo tiene un codo en $d = 2$, es por este motivo que el algoritmo en todos los casos elige dos variables. Además ninguno de estos conjuntos contiene sólo información de una sola de las variables originales. También podemos observar en la Figura 4.3 que la curva da muy similar eligiendo la cantidad de vecinos por cross–validación o fijando una cantidad arbitraria, es decir, el procedimiento no es sensible a la cantidad de vecinos considerada para estimar la esperanza condicional.

Es importante notar que Zou et al. (2006) eligen las cuatro primeras variables para la primera componente principal y las variables X_{ij} con $i = 5, \ldots, 8$, para la segunda componente. Es decir, que su procedimiento de selección de variables detecta información repetida.

Para hacer el problema más complejo repetimos el experimento considerando $\sigma^2 = 100$. En este caso, las dos primeras componentes principales explican más de un 75% de la varianza. Nuevamente aplicamos las dos propuestas y consideramos, para cada una de ellas, los mismos parámetros que para el problema anterior.

Primero aplicamos el Procedimiento *Blinding* CP1. En la Figura 4.2 observamos que al considerar un ángulo de 10 o 15 grados la cantidad de variables seleccionadas es mayor a dos, situación que se revierte al considerar un ángulo de 20 o 25 grados donde elige en la mayoría de los casos dos variables. En los casos que elige dos variables nunca escoge un conjunto que sólo dependa de una de las variables originales.

Luego aplicamos el Procedimiento *Blinding* M. En la Figura 4.3 vemos que el valor óptimo de la función objetivo tiene un codo en d=2, eligiendo por tal motivo dos variables. A su vez, en todos los casos las variables tienen información sobre las dos variables originales. Podemos observar en el gráfico que la curva es muy similar si, para estimar la esperanza condicional, fijamos la cantidad de vecinos o los encontramos por cross–validación, obteniendo conclusiones análogas.

A continuación analizamos el desempeño del Procedimiento *Blinding* M en presencia de dependencia no lineal.

Sean

$$X_{ij} = \begin{cases} Y_{1j} + e_{ij} & \text{para } i = 1, 2, \\ Y_{2j} + e_{ij} & \text{para } i = 3, 4, \\ Y_{3j} + e_{ij} & \text{para } i = 5, \\ \sqrt{|Y_{1j} Y_{2j}|} + e_{ij} & \text{para } i = 6, \\ \log Y_{1j}^2 + e_{ij} & \text{para } i = 7, \\ 12 \log Y_{2j}^2 + e_{ij} & \text{para } i = 8, \end{cases} \tag{4.9}$$

donde los errores e_{ij} siguen la misma distribución que en los ejemplos anteriores. La proporción de la varianza explicada prácticamente en todos los casos estuvo entre el 70% y el 80% (respec. 60% y el 70%) para $\sigma^2 = 1$ (respec. $\sigma^2 = 100$). Realizamos 1000 réplicas. El número de vecinos más cercanos para estimar la esperanza condicional fue hallado, al igual que para los problemas previos, por cross–validación y también fijando 5, 10 y 20 vecinos. En todos los casos los resultados son muy similares. El algoritmo elige dos variables en más de un 93% de las réplicas al fijar la cantidad de vecinos y en más de un 97% de los casos al cross–validar la cantidad de vecinos y en los casos restantes siempre elige 3 variables. En la Figura 4.4 observamos cuanto vale en cada una de las réplicas el valor óptimo de la función objetivo para 1,2,... y 8 variables, también se puede observar la curva media. Se ve claramente que en la mayoría de los casos el número óptimo de variables es 2. Solamente en un 11% (respec. 32%) de los casos para $\sigma^2 = 1$ (respec. $\sigma^2 = 100$) elige un conjunto de variables que sólo dependa de una de las variables originales.

Comparamos el Procedimiento *Blinding* M con otros procedimientos de selección de variables. Algunos métodos estiman con cuantas variables quedarse mientras que a otros el usuario le tiene que indicar el número de variables. El Procedimiento *Blinding* M, al igual que los de componentes principales esparzas corresponden a los de la primera categoría. Un ejemplo de componentes esparzas es el de SPC vía PMD (Witten et al. (2009)). Entre los métodos que corresponden a la otra categoría encontramos los propuestos por Jolliffe (2002), como por ejemplo J1 o J2 y el algoritmo propuesto por McCabe (1984).

Es más justo comparar nuestro método con el de SPCA vía PMD. Por este motivo, aplicamos este método al modelo (4.9) y observamos que siempre seleccionan todas las variables, y en más del 80% de los casos 4 variables tienen un alto peso en las primeras dos componentes, es decir, SPCA elige información redundante. Si analizamos el mismo ejemplo para J1, J2 y McCabe, seleccionando dos variables obtenemos resultados similares en los tres casos. Las dos variables elegidas que no retienen la información de Y_1 o de Y_2 entre un 40% y un 63% (respec. entre un 43% y un 47%) de las veces para $\sigma^2 = 1$ (respec. $\sigma^2 = 100$). En todos los casos el Procedimiento *Blinding* M realiza mejores elecciones.

Finalmente, agregamos al modelo (4.9) 60 variables ruido independientes normalmente distribuidas con esperanza cero y varianza $\sigma^2 = 1$ ó 100. En cada caso realizamos 200 réplicas e hicimos una búsqueda exhaustiva de hasta tres variables. Al realizar la estimación no paramétrica de la esperanza condicional no consideramos el caso de búsqueda exhaustiva para la cantidad de vecinos. En todos los casos dos variables fueron elegidas y nunca se eligieron variables ruido, solo en un 10% (respec. 30%) de los casos, para $\sigma^2 = 1$ (respec.

$\sigma^2 = 100$) las variables elegidas fueron inadecuadas.

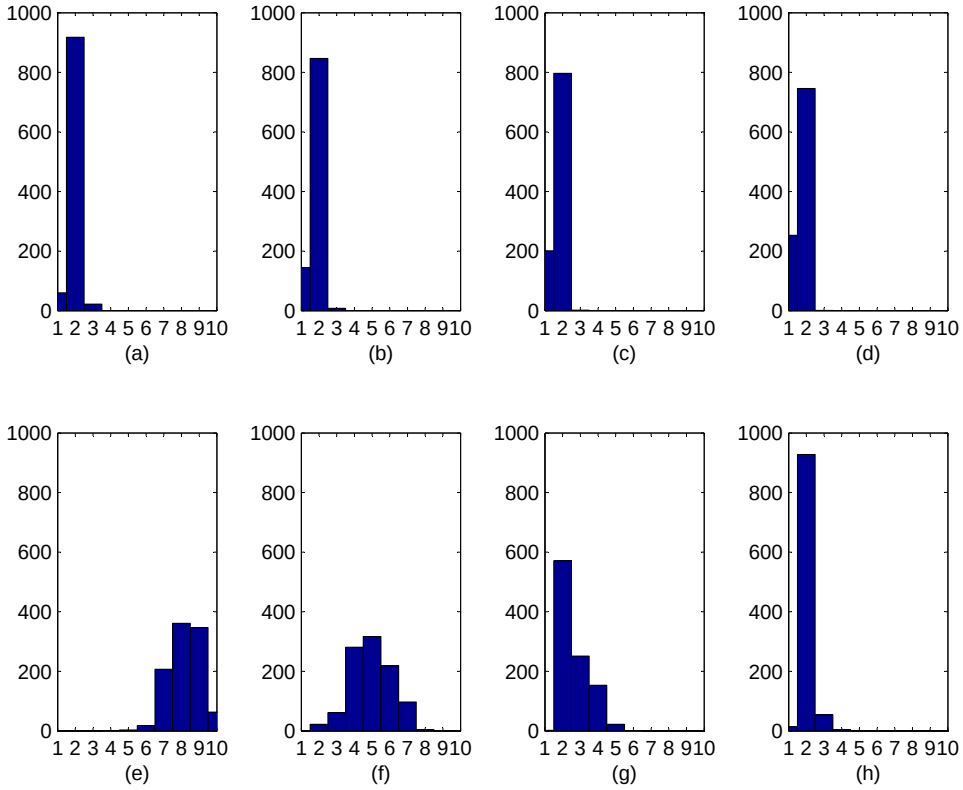

Figura 4.2: Los Histogramas indican el número de veces que se eligen d variables mediante el Procedimiento *Blinding* CP1. (a) $\sigma^2 = 1$, ángulo = 10 grados (b) $\sigma^2 = 1$, ángulo = 15 grados, (c) $\sigma^2 = 1$, ángulo = 20 grados, (d) $\sigma^2 = 1$, ángulo = 25 grados, (e) $\sigma^2 = 100$, ángulo = 10 grados, (f) $\sigma^2 = 100$, ángulo = 15 grados, (g) $\sigma^2 = 100$, ángulo = 20 grados, (h) $\sigma^2 = 100$, ángulo = 25 grados.

4.4 Ejemplos Reales

4.4.1 Regresión: Diabetes Data Set

Efron et al. (2004) analizaron los datos de pacientes con diabetes para estudiar el desempeño de su algoritmo LARS. El conjunto de datos consta de 10 variables predictoras (edad, sexo, masa corporal, presión arterial media y seis mediciones en suero sanguíneo) medidas en 442 pacientes diabéticos y una variable de respuesta, la glucemia, que permite medir la progresión de la enfermedad luego de un año.

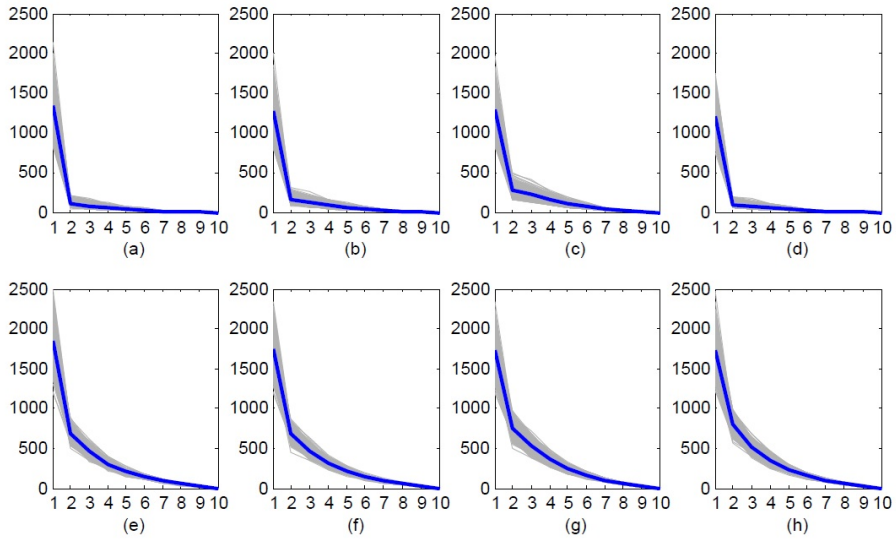

Figura 4.3: Valor óptimo de la función objetivo en cada réplica, seleccionando variables mediante el Procedimiento *Blinding* M para el caso lineal. La línea azul indica la curva media. (a) $\sigma^2 = 1$, $r = 5$, (b) $\sigma^2 = 1$, $r = 10$, (c) $\sigma^2 = 1$, $r = 20$, (d) $\sigma^2 = 1$, r=búsqueda exhaustiva, (e) $\sigma^2 = 100$, $r = 5$, (f) $\sigma^2 = 100$, $r = 10$, (g) $\sigma^2 = 100$, $r = 20$, (h) $\sigma^2 = 100$, r=búsqueda exhaustiva.

Efron et al. (2004) aplican LARS a este conjunto de datos e identifican que para predecir la evolución de la enfermedad, sólo 4 de estas variables son relevantes: la masa corporal, la presión arterial media y dos de las mediciones al suero sanguíneo (variable 7 y 9). Los autores sugieren quedarse con estas variables y luego ajustan un modelo clásico de regresión lineal.

Nuestro objetivo es determinar si las cuatro variables son necesarias para interpretar el modelo o si hay información redundante. Como el número de variables es pequeño, hicimos un análisis exhaustivo sobre todos los posibles subconjuntos de variables.

En la Figura 4.5(a) vemos para los diferentes cardinales de los conjuntos de variables los valores óptimos de la función objetivo. Notamos que la función tiene un codo pronunciado en dos variables. El subconjunto de variables de cardinal 2 que minimiza la función objetivo (4.2) está conformado por las variables masa corporal y una de las mediciones en suero sanguíneo (variable 9). El número de vecinos más cercanos en la estimación de la esperanza condicional fue cross–validado y resultó ser 86 en presión arterial y 107 para la variable 7.

En la Figura 4.5(b) mostramos los valores que se obtienen al aplicar la regresión con las variables originales y con las variables *blinded*. Si no hubiera ningún error entonces los puntos deberían de encontrarse sobre la recta identidad, podemos ver que la estimación es muy buena.

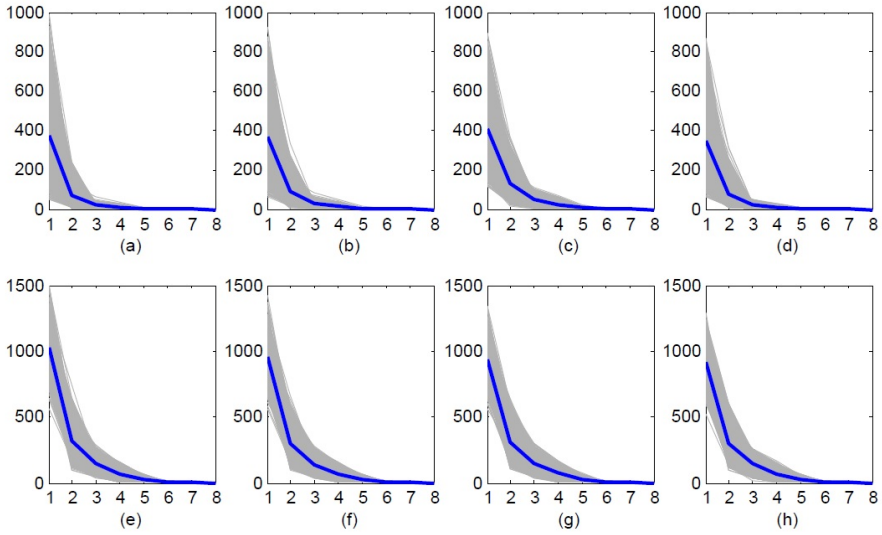

Figura 4.4: Valor óptimo de la función objetivo en cada réplica, seleccionando variables mediante el Procedimiento *Blinding* M para el caso no lineal. La línea azul indica la curva media. (a) $\sigma^2 = 1$, $r = 5$, (b) $\sigma^2 = 1$, $r = 10$, (c) $\sigma^2 = 1$, $r = 20$, (d) $\sigma^2 = 1$, r=búsqueda exhaustiva, (e) $\sigma^2 = 100$, $r = 5$, (f) $\sigma^2 = 100$, $r = 10$, (g) $\sigma^2 = 100$, $r = 20$, (h) $\sigma^2 = 100$, r=búsqueda exhaustiva.

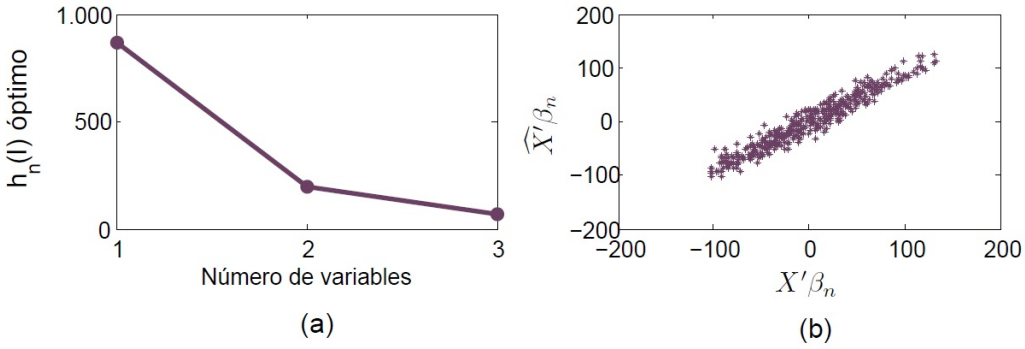

Figura 4.5: (a) Valores óptimos de la función objetivo en función del número de variables. (b) Scatter plot de los valores predichos considerando las observaciones *blinded* utilizando el subconjunto óptimo versus los valores predichos con las observaciones originales.

4.4.2 Modelo Lineal Generalizado: South African Heart Disease Data Set

En Sudáfrica se realizó un estudio para analizar los factores de riesgo de la enfermedad isquémica cardíaca, CORIS(Coronary Risk–Factor Study) (ver Rousseauw et al. (1983)).

Hastie et al. (2001) analizaron un subconjunto de estos datos que pertenecen a la encuesta realizada en tres áreas rurales de la provincia Cabo Occidental de Sudáfrica. El objetivo del estudio era establecer la prevalencia de los diferentes factores de riesgo de la enfermedad en esa región. Los datos cuentan con 7 variables predictoras y una de respuesta que indica la presencia o ausencia del infarto de miocardio al momento del estudio, estas variables fueron medidas en 462 hombres blancos entre 15 y 64 años.

Realizaron dos procedimientos backward de selección de variables para modelos lineales generalizados. Por un lado, eliminan en cada paso la variable menos informativa. Por otro lado, para decidir que variable eliminar, comparan los diferentes modelos que se obtienen al quitar una variable mediante un análisis de la deviance. En ambos casos concluyen que cuatro variables son necesarias: edad del paciente, cantidad de tabaco consumido, concentración de colesterol y de lipoproteínas de baja densidad e historia familiar de ataques cardíacos. Las primeras tres variables son continuas, pero la última es binaria.

Analizamos este conjunto de datos usando el Procedimiento *Blinding* M de selección de variables. La Figura 4.6(a) indica que es conveniente elegir dos variables. Estas son, historia familiar y edad del paciente, notemos que estas variables también fueron elegidas por Hastie et al. (2001). En la Figura 4.6(b) observamos que la predicción lineal usando las variables originales y usando las variables *blinded* está muy correlacionada. Al estimar en forma no paramétrica la esperanza condicional, la cantidad de vecinos más cercanos osciló entre 72 y 150 en los diferentes casos.

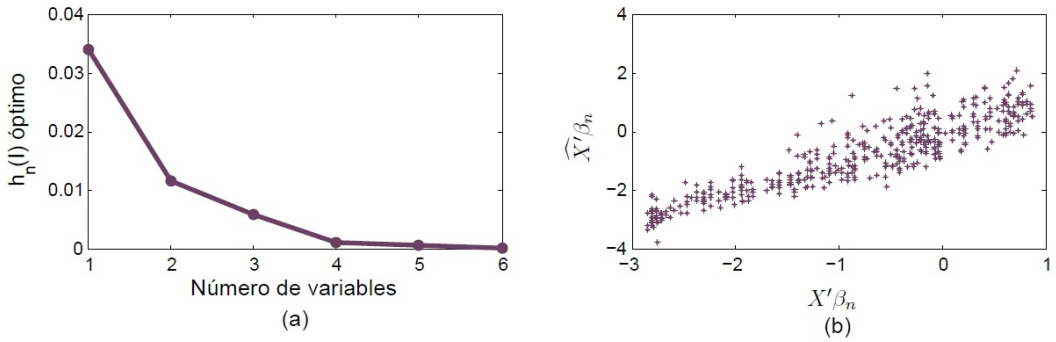

Figura 4.6: (a) Valores óptimos de la función objetivo en función del número de variables. (b) Scatter plot de la predicción lineal luego de aplicar el Procedimiento *Blinding* M versus la predicción lineal considerando los datos originales.

4.4.3 Componentes Principales: Alate Adelges Data Set

En 1964 mediante una trampa lumínica se capturaron pulgones alados. Uno de los objetivos del estudio era determinar el número de subespecies que estaban presentes en ese hábitat. Los datos eran muy difíciles de clasificar usando métodos convencionales. El primero en analizar

los datos fue Jeffers (1967) quien disponía de 19 medidas de 40 pulgones alados. Como la naturaleza y escala de las variables era muy diferente realizó un análisis de componentes normadas, es decir, utilizó la matriz de correlación para hacer el análisis de componentes principales. Las dos primeras componentes explican el 86% de la variabilidad de los datos. Jeffers (1967) observó que cuando uno proyectaba los datos en las dos primeras componentes principales se podían observar claramente cuatro grupos. La primera componente puede ser vista como un promedio de las medidas de tamaño, mientras que la segunda componente está dominada por la mayoría de las variables discretas.

Jolliffe (2002) hizo un estudio de selección de variables para estos datos. Consideró para el análisis de componentes principales tres procedimientos heurísticos diferentes, sus métodos J1 y J2, McCabe y un procedimiento propuesto por Cadima & Jolliffe (2001) que se basa en buscar el subconjunto de variables cuyo espacio generado aproxime mejor al espacio generado por las primeras l componentes principales. Todos ellos sugieren quedarse con tres o cuatro variables. Jolliffe (2002) señaló que si uno se queda con dos componentes principales, entonces es deseable elegir dos o poco más de dos variables.

En nuestro caso, la Figura 4.7(a) nos sugiere quedarnos con dos variables. Las variables que elige nuestro procedimiento son la longitud del ala trasera (variable 4) y la longitud del ovipositor (variable 16). La primera variable nos da una idea del tamaño del insecto y tiene mucho peso en la primera componente, la segunda tiene mucho peso en la segunda componente y su peso tampoco es despreciable en la primera componente. Observamos en la Figura 4.7(b) que si nos quedamos con estas dos variables y estimamos a las otras con la esperanza condicional, cuando proyectamos en las dos primeras componentes los cuatro grupos también son claramente distinguibles. Si nos quedamos con una sola variable la estructura de los cuatro grupos no se distingue fácilmente. La cantidad de vecinos para cada variable fue cross–validada seleccionando entre 2 y 20 vecinos para cada variable predicha, siendo en la mayoría de los casos inferior a 10 vecinos.

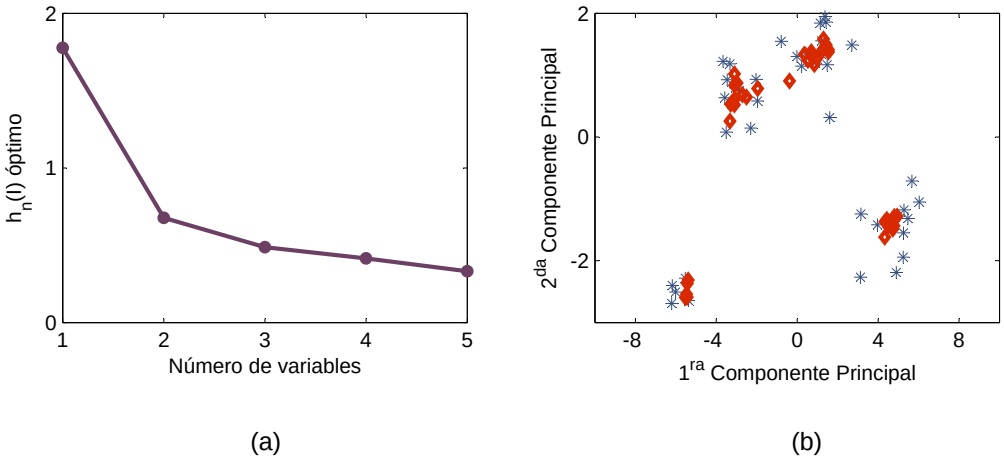

Figura 4.7: (a) Valores óptimos de la función objetivo en función del número de variables. (b) Scatter plot de los pesos de las dos primeras componentes principales considerando las variables originales (estrellas azules) y considerando las variables estimadas (diamantes rojos).

Capítulo 5

Selección de Variables para Datos Funcionales

Los problemas estadísticos estudiados en los capítulos anteriores tienen su correlato en el contexto de datos funcionales. Al variar la formulación y solución de los problemas estadísticos es necesario pensar estrategias nuevas para el problema de selección de variables. En este capítulo abordamos esta cuestión para los casos de componentes principales, regresión y clasificación funcional. Damos los resultados de consistencia correspondientes.

5.1 Introducción

Los avances tecnológicos de los últimos años posibilitaron procesar la información en tiempo real, esto provocó que en diferentes áreas, en forma asidua, se utilicen cada vez más modelos que involucran datos funcionales. Esto sucede, por ejemplo, con los precios de las acciones, en meteorología hay fenómenos espacio temporales como la tormenta del Niño o en las imágenes satelitales, diferentes estudios médicos son evaluados en tiempo real usando electrocardiogramas o electroencefalogramas. En este contexto de amplia presencia de datos funcionales fue necesario estudiar con este enfoque los problemas clásicos de la estadística.

En este camino, nuevos problemas han surgido, por tal motivo hay una vasta literatura enfocada a extender estos métodos tanto desde un punto de vista teórico como empírico.

Uno de los primeros modelos estudiados para datos funcionales fue el de componentes principales. La primera referencia del tema fue dada por Grenander (1950). Sin embargo, se comenzó a estudiar de manera sistemática a principios de la década pasada. En los libros de Ramsay & Silverman (2005), Ferraty & Romain (2011) y Horváth & Kokoszka (2012) hay capítulos dedicados a estos temas donde se encuentran referencias específicas de manera detallada.

Otros problemas muy estudiados son los modelos de regresión lineal. Dentro de estos modelos hay dos grandes categorías aquellos cuya respuesta es escalar y otros cuya respuesta

es funcional. Estos últimos fueron introducidos por Ramsay & Dalzell (1991). En los libros de Ramsay & Silverman (2005), Ferraty & Romain (2011) y Horváth & Kokoszka (2012) hay capítulos dedicados a estos modelos donde dan un amplio panorama del tema.

También ha sido abordado el tema de clasificación de datos funcionales, en el libro de Ferraty & Romain (2011) dan un panorama general del tema. Tian & James (2013) dan una propuesta que se aproxima más al enfoque que nosotros encaramos por tal motivo explicamos con mayor detalle su propuesta. En lugar de resolver el problema en un espacio de dimensión infinita, buscan reducir la dimensión del espacio y luego clasificar en un espacio de dimensión baja. Sea $\{X(t),\ t \in [a,b]\}$ un proceso estocástico y una respuesta categórica Y, sin pérdida de generalidad $Y = \{0,1\}$. Se consideran las proyecciones de $X(t)$ respecto de las funciones $\left\{f_1,\ldots,f_p\right\}$ en $[a,b]$,

$$Z_j = \int_a^b X(t) f_j(t) dt \text{ con } \ j = 1,\ldots,p.$$

La propuesta se basa en elegir las funciones $\left\{f_1,\ldots,f_p\right\}$ tales que minimicen el error de clasificación al resolver el problema en dimensión baja en lugar del espacio original con la restricción de que las funciones tomen formas simples. Ellos consideran que pertenezcan al grupo de funciones que son constantes o lineales (creciente o decreciente) en un intervalo $[a_1,b_1] \subset [a,b]$ y nulas fuera de él. En su trabajo también proponen un algoritmo el cual mediante estudios de simulación demuestran que es competitivo para resolver este problema que es computacionalmente costoso.

Nuestro enfoque es diferente. Supongamos que llevamos a cabo un procedimiento estadístico (clasificación, regresión, componentes principales, etc) en forma exitosa para un conjunto de datos funcionales y que además contamos con un conjunto de funciones $\left\{f_1,\ldots,f_p\right\}$ mediante las cuales se puede describir características "importantes" de estos datos. El objetivo es explicar del mejor modo posible, mediante un subconjunto de funciones de $\left\{f_1,\ldots,f_p\right\}$, el procedimiento estadístico en cuestión. Para realizarlo extendemos la idea de *Blinding*. Este método al igual que los anteriores es ad hoc al modelo estadístico.

5.2 Procedimiento *Blinding* Funcional

Sea $X : \Omega \to L^2[a,b]$ un elemento aleatorio, $\{X = X(t) : t \in [a,b]\}$ en un espacio de probabilidad $(\Omega, \mathcal{A}, P)$ donde $[a,b] \subset \mathbb{R}$ es un intervalo finito. Sea $\mathbb{G}$ un conjunto de acciones y $\mathbb{E}$ un espacio métrico separable, que en general consideraremos $L^2[a,b]$. Un *modelo estadístico poblacional* está dado por una función $\psi(P) = \psi(X,g)$, donde $\psi : L^2[a,b] \times \mathbb{G} \to \mathbb{E}$. La salida del procedimiento estadístico es un elemento aleatorio en $\mathbb{E}$ dada por $\psi(X,g)$, típicamente una función aleatoria en $L^2[a,b]$.

Dado un conjunto de funciones conocidas, de las trayectorias, a valores reales,

$$\left\{f_1,\ldots,f_p,\ f_i : \left(L^2[a,b],P\right) \to \mathbb{R} \text{ para todo } 1 \le i \le p\right\},$$

el objetivo es seleccionar un pequeño subconjunto de ellas que describa del mejor modo posible el resultado obtenido al aplicar el método estadístico.

Algunos ejemplos de conjuntos de funciones en los que vamos a estar interesados son:

(a) Evaluaciones puntuales:

$f_1(X) = X(t_1), \ldots, f_p(X) = X(t_p)$ para $a \leq t_1 < t_2 < \ldots < t_p \leq b$.

(b) Promedios locales:

$f_1(X) = \frac{1}{|T_1|}\int_{T_1} X(t)dt, \ldots, f_p(X) = \frac{1}{|T_p|}\int_{T_p} X(t)dt$ donde $\left\{T_1, \ldots, T_p\right\}$ son subintervalos disjuntos del intervalo $[a, b]$.

(c) Medida de ocupación de un conjunto:

$f_1(X) = |\{t : X(t) \in T_1\}|, \ldots, f_p(X) = \left|\left\{t : X(t) \in T_p\right\}\right|$, donde $T_1, \ldots, T_p$ son intervalos disjuntos (no necesariamente acotados) en $\mathbb{R}$ y denotamos por $|.|$ a la medida de Lebesgue del conjunto. Es decir, estamos considerando la *medida de ocupación* de los subconjuntos T_j, es decir cuanto tiempo nuestro proceso permanece en cierto intervalo o cuanto tiempo se encuentra por encima o debajo de cierto nivel.

(d) Cruces de nivel:

$f_i(X) =$ número de cruces ascendentes a cierto nivel t_i con $i = 1, \ldots, p$ para $t_1 < t_2 < \ldots < t_p \in \mathbb{R}$.

(e) Momentos de las normas:

$f_1(X) = E\left(||X||_{L^2[a,b]}\right), \ldots, f_p(X) = E\left(||X||^p_{L^2[a,b]}\right)$.

(f) Momentos:

$f_1(X) = \int_a^b X(t,\omega)dP(\omega), \ldots, f_p(X) = \int_a^b X^p(t,\omega)dP(\omega)$.

Llamamos

$$\mathbf{f}(X) = \mathbf{f} = \left(f_1, \ldots, f_p\right) = \left(f_1(X), \ldots, f_p(X)\right),$$

y dado un conjunto de índices $I = \{i_1 < \ldots < i_d\} \subset \{1, \ldots, p\}$ con $d \leq p$, consideramos el vector aleatorio

$$\mathbf{f}(I, X) = \mathbf{f}(I) = (f_{i_1}(X), \ldots, f_{i_d}(X)) \in \mathbb{R}^d.$$

Definición 2 *Sea I un subconjunto de $\{1, \ldots, p\}$ llamamos **proceso estocástico blinded** de X respecto del vector aleatorio $\mathbf{f}(I)$ al proceso $Z(I) : [a, b] \to \mathbb{R}$, tal que*

$$Z(I)(t) = E(X(t)|\mathbf{f}(I)) = \eta(t, \mathbf{f}(I, X)). \tag{5.1}$$

Notamos $Q(I)$ a la distribución del proceso $Z(I)$.

Observación 7 *$Z(I)(t)$ es un proceso estocástico, aunque la esperanza condicional está tomada respecto de un vector aleatorio finito dimensional, $\mathbf{f}(I) \in \mathbb{R}^d$.*

Siguiendo las mismas ideas de los capítulos anteriores describiremos el Procedimiento *Blinding* Funcional (F). Para cada entero fijo d, $1 \leq d \leq p$ consideramos $\mathcal{I}_d$ la familia de todos los subconjuntos de cardinal d y buscamos el conjunto $I \in \mathcal{I}_d$ tal que $\psi(X, g)$ este lo más cerca posible de $\psi(Z(I), g)$. La noción de cercanía va a variar en cada uno de los diferentes modelos y la definimos mediante una función $h(I, P, Q(I), \psi) = h(I)$.

Más precisamente, definimos $\mathcal{I}_0 \subset \mathcal{I}_d$ la familia de todos los subconjuntos para los cuales el mínimo de la función $h(I)$ es alcanzado, es decir,

$$\mathcal{I}_0 = arg \min_{I \in \mathcal{I}_d} h(I). \tag{5.2}$$

Para describir la versión empírica es necesario estimar en forma consistente al conjunto I_0, $I_0 \subseteq \mathcal{I}_0$ basado en la muestra de trayectorias del proceso estocástico X, $\{X_1, \ldots, X_n\}$.

Dado un conjunto $I \in \mathcal{I}_d$, el primer paso es obtener la versión *blinded* de la muestra aleatoria, es decir $\left\{\hat{X}_1(I), \ldots, \hat{X}_n(I)\right\}$, que solo dependa de $\mathbf{f}(I, X)$, estimando la esperanza condicional mediante un estimador no paramétrico. A la distribución de los elementos aleatorios $\left\{\hat{X}_j(I), 1 \leq j \leq n\right\}$ la denotaremos por $Q_n(I)$.

A modo de ejemplo, nosotros consideramos el estimador de vecinos más cercanos (r-NN). Se fija un entero r (el número de vecinos más cercanos a considerar) y para cada $j \in \{1, \ldots, n\}$ se busca un conjunto de índices C_j de los r vecinos más cercanos de $\mathbf{f}(I, X_j)$ entre $\{\mathbf{f}(I, X_1), \ldots, \mathbf{f}(I, X_n)\}$. Luego definimos los elementos aleatorios

$$\hat{X}_j(I)(t) = \hat{E}\left(X_j(t)|\mathbf{f}(I)\right) = \frac{1}{r} \sum_{m \in C_j} X_m(t) \in L^2[a, b].$$

Observemos que el conjunto C_j no depende de t.

De este modo, para cada conjunto $I \in \mathcal{I}_d$, definimos la versión empírica de la función objetivo $h_n(I)$, y buscamos los subconjuntos que minimicen dicha función, es decir,

$$\mathcal{I}_n = arg \min_{I \in \mathcal{I}_d} h_n(I). \tag{5.3}$$

Para probar la consistencia del criterio propuesto en los diferentes modelos, necesitamos la siguiente hipótesis:

Hf1. $\hat{X}(I)$ es un estimador fuertemente consistente de $Z(I)$, es decir,

$$\left\|Z(I) - \hat{X}(I)\right\|_{L^2[a,b]} \to_{c.s.} 0.$$

Las condiciones bajo las cuales se cumple **Hf1** son un caso particular del trabajo de Lian (2011). En el mismo se establecen condiciones más generales para obtener consistencia ya que se trabaja en un espacio con una pseudométrica. En nuestro caso las variables están en $\mathbb{R}^d$, por eso podemos relajar algunas restricciones. Para ser más precisos, tenemos el siguiente resultado:

Proposición 1 *Sea $X(t) = \eta(t, \mathbf{f}(I,X)) + e$, $X(t) \in \mathcal{H}$, un espacio de Hilbert separable y $\mathbf{f}(I,X) \in \mathbb{R}^d$, $e \in \mathcal{H}$ con esperanza cero e independiente de $\mathbf{f}(I,X)$. Sea η_n el estimador no paramétrico de la esperanza condicional dado por vecinos más cercanos (r-NN). Bajo las siguientes hipótesis,*

1. *$\mathbf{f}(I,X)$ tiene una función de densidad g que cumple $0 < c_1 \leq g(x) \leq c_2$ para todo x del soporte de $\mathbf{f}(I,X)$,*

2. *(a) $\|\eta(t, \mathbf{f}(I,X))\|_{\mathcal{H}} \leq B$ para todo $\mathbf{f}(I,X) \in \mathbb{R}^d$,*

   (b) $\|\eta(t, \mathbf{f}(I,X)) - \eta(t, \mathbf{f}(I,X)')\|_{\mathcal{H}} \leq M \|\mathbf{f}(I,X) - \mathbf{f}(I,X)'\|_2$ (la condición de Lipschitz),

3. *existe $\delta > 0$ tal que $E\left(\|e\|_{\mathcal{H}}^{2+\delta}\right) < \infty$,*

4. *$k/n \to 0$, $k/logn \to \infty$, $\sum_{n=1}^{\infty} (logn/k)^{\delta/2} < \infty$,*

tenemos que

$$\|\eta_n(t, \mathbf{f}(I,X)) - \eta(t, \mathbf{f}(I,X))\|_{\mathcal{H}} = O\left(\left(\frac{k}{n}\right)^{1/d} + \sqrt{\frac{logn}{k}}\right) \quad c.s.$$

Observación 8 *Esta Proposición es una consecuencia directa del Teorema 1 de Lian (2011) y el hecho que en $\mathbb{R}^d$ bajo la condición 1 por el Teorema de Diferenciación de Lebesgue se tiene que $P(B(x,h)) = O\left(h^d\right)$. Notemos también que en nuestro caso el espacio de Hilbert es $L^2[a,b]$.*

A continuación definimos en forma explícita las funciones objetivo teóricas y empíricas para los diferentes problemas estudiados.

5.2.1 Clasificación supervisada y no supervisada

Sea $\left\{X(t) \in L^2[a,b]\right\}$ un proceso estocástico y K el número de clusters. Un procedimiento de cluster está determinado por una función $g : L^2[a,b] \to \{1, \ldots, K\}$ que asigna a que grupo pertenece la trayectoria. Denotamos con $G_k = g^{-1}(k)$ con $k = 1, \ldots, K$ a la partición del espacio.

En este caso, la función objetivo es

$$h(I) = 1 - \sum_{k=1}^{K} P(g(X) = k, g(Z(I)) = k). \tag{5.4}$$

Esta función mide la diferencia entre la partición del espacio considerando la trayectoria original y la *blinded*. Luego, dado $d < p$, buscamos un conjunto $I \in \mathcal{I}_d$ que minimice la función objetivo (5.4), como mencionamos en (5.2).

Observación 9 *Este criterio es análogo a buscar los conjuntos que maximicen la función,*

$$h^*(I) = \sum_{k=1}^{K} P(g(X) = k, g(Z(I)) = k),$$

que es la función objetivo definida por Fraiman et al. (2008) para el caso multivariado.

La versión empírica de la función (5.4) está dada por,

$$h_n(I) = 1 - \frac{1}{n} \sum_{k=1}^{K} \sum_{j=1}^{n} I_{\{g_n(X_j)=k\}} I_{\{g_n(\hat{X}_j(I))=k\}}, \tag{5.5}$$

donde la función $g_n : L^2[a,b] \to \{1,\dots,K\}$ determina a que grupo pertenece cada proceso empírico. De este modo, la partición del espacio está dada por $G_k^n = g_n^{-1}(k)$ con $k = 1,\dots,K$. El objetivo es hallar los conjuntos de funciones que minimicen la ecuación (5.5), del mismo modo que en (5.3).

Para establecer el resultado de consistencia, además de necesitar la condición **Hf1** definida anteriormente pedimos las siguientes hipótesis adicionales:

HfC1. (a) La partición del espacio es fuertemente consistente, es decir, dado $\epsilon > 0$, existe un conjunto $A(\epsilon) \subset L^2[a,b]$ con $P(X \in A(\epsilon)) > 1 - \epsilon$ tal que, para todo $r > 0$, $sup_{x \in C(\epsilon,r)} \left| I_{\{g_n(x)=k\}} - I_{\{g(x)=k\}} \right| \to_{c.s.} 0$ cuando $n \to \infty$ para $k = 1,\dots,K$, donde $C(\epsilon, r) = A(\epsilon) \cap \mathcal{K}_r$ con $\mathcal{K}_r$ una secuencia creciente de conjuntos compactos tal que $P(X \in \mathcal{K}_r) \to 1$ cuando $r \to \infty$.

(b) $d(X, \partial G_k^n) - d(X, \partial G_k) \to_{c.s.} 0$ cuando $n \to \infty$, donde ∂G_k (respec. ∂G_k^n) es la frontera del conjunto G_k (respec. G_k^n).

HfC2. $P(d(Z(I), \partial G_k) < \delta) \to 0$ cuando $\delta \to 0$ para $k = 1,\dots,K$.

HfC3. La distribución es no degenerada, es decir, para todo $\delta > 0$, $P(X \in B(x,\delta)) > 0$ para casi todo $x \in L^2[a,b]$.

Teorema 6 *Sean* $\left\{X_j(t) : t \in [a,b]\right\}$ *realizaciones independientes e idénticamente distribuidas del proceso estocástico con la misma distribución que* $X(t)$*. Dado* d, $1 \leq d \leq p$*, sea* $\mathcal{I}_d$ *la familia de todos los conjuntos de* $\{1,\ldots,p\}$ *con cardinal* d *y sea* $\mathcal{I}_0 \subset \mathcal{I}_d$ *la familia de todos los conjuntos para los cuales se alcanza el mínimo de la función (5.4). Bajo* **Hf1**, **HfC1**, **HfC2** *y* **HfC3** *tenemos que para cada* $I_n \in \mathcal{I}_n$*, existe un* $n_0(\omega)$ *tal que para todo* $n > n_0(\omega)$*, con probabilidad uno* $I_n \in \mathcal{I}_0$*.*

La demostración del Teorema 6 se encuentra en el último capítulo.

5.2.2 Componentes Principales

Sea $\left\{X(t) \in L^2[a,b]\right\}$ un proceso estocástico con trayectorias continuas, que sin pérdida de generalidad asumiremos que tiene esperanza cero para casi todo $t \in [a,b]$, es decir $E\left(X(t)\right) = 0$ y pediremos que tenga segundo momento finito, es decir $E\left(\|X\|^2_{L^2[a,b]}\right) < \infty$.

Definimos,

$$v(t,s) = E(X(t)X(s)),$$

su función de covarianza que tiene asociado un operador lineal $\mathbf{\Gamma} : L^2[a,b] \to L^2[a,b]$ definido como,

$$\left(\mathbf{\Gamma}u\right)(t) = \int_a^b v(t,s)u(s)ds \text{ para todo } u \in L^2[a,b],\ t \in [a,b]. \tag{5.6}$$

Asumimos que

$$\|v\|^2_{L^2([a,b]\times[a,b])} = \int_a^b \int_a^b v^2(t,s)dtds < \infty. \tag{5.7}$$

A continuación veamos algunas propiedades del operador de covarianza $\mathbf{\Gamma}$.

(a) Linealidad. Sean $u, v \in L^2[a,b]$, $c \in \mathbb{R}$ y $t \in [a,b]$,

$$\mathbf{\Gamma}(cu+v)(t) = \int_a^b v(t,s)(cu+v)(s)ds = c\int_a^b v(t,s)u(s)ds + \int_a^b v(t,s)v(s)ds = c\mathbf{\Gamma}u(t) + \mathbf{\Gamma}v(t).$$

(b) Continuidad. Como consecuencia de la desigualdad de Cauchy-Schwarz obtenemos

$$\begin{aligned}
\|\mathbf{\Gamma}u\|^2_{L^2[a,b]} &= \int_a^b \left(\int_a^b v(t,s)u(s)ds\right)^2 dt = \int_a^b \langle v(t,\cdot), u\rangle^2_{L^2[a,b]}\, dt \\
&\leq \int_a^b \|v(t,\cdot)\|^2_{L^2[a,b]}\, \|u\|^2_{L^2[a,b]} dt = \int_a^b \int_a^b v^2(t,s)ds \int_a^b u^2(s)\, dsdt \\
&= \int_a^b \int_a^b v^2(t,s)dsdt \int_a^b u^2(s)\, ds = \|v\|^2_{L^2([a,b]\times[a,b])}\|u\|^2_{L^2[a,b]},
\end{aligned}$$

y por (5.7) tenemos, $\|\mathbf{\Gamma}u\|_{L^2[a,b]} \leq K\|u\|_{L^2[a,b]}$ con $K > 0$.

(c) Acotado. Es una consecuencia inmediata de la continuidad, ya que

$$\max_{\|u\|_{L^2[a,b]}=1} \|\mathbf{\Gamma} u\|_{L^2[a,b]} \leq K \text{ con } K \in \mathbb{R}_{>0}.$$

(d) Autoadjunto. Sean $u, v \in L^2[a,b]$, tenemos que

$$\langle u, \mathbf{\Gamma} v\rangle_{L^2[a,b]} = \int_a^b u(t) \int_a^b v(t,s)v(s)dsdt = \int_a^b v(s) \int_a^b v(t,s)u(t)dtds = \langle \mathbf{\Gamma} u, v\rangle_{L^2[a,b]}.$$

Es decir, $\mathbf{\Gamma}$ es un operador lineal, continuo, acotado y autoadjunto.

Denotemos con $\mathcal{F}$ al espacio de Hilbert de dicho operador, luego su producto interno está definido por:

$$\langle \mathbf{\Gamma_1}, \mathbf{\Gamma_2}\rangle_{\mathcal{F}} = \text{traza}\,(\mathbf{\Gamma_1}\mathbf{\Gamma_2}) = \sum_{j=1}^{\infty} \left\langle \mathbf{\Gamma_1} u_j, \mathbf{\Gamma_2} u_j \right\rangle_{L^2[a,b]},$$

donde $\left\{u_j, j \geq 1\right\}$ es una base ortonormal de $L^2[a,b]$.

Considerando el Teorema de Schmidt se tiene que,

$$\begin{aligned}
\|\mathbf{\Gamma}\|^2_{\mathcal{F}} = \langle \mathbf{\Gamma}, \mathbf{\Gamma}\rangle_{\mathcal{F}} = \sum_{j=1}^{\infty} \left\langle \mathbf{\Gamma} u_j, \mathbf{\Gamma} u_j \right\rangle_{L^2[a,b]} &= \sum_{j=1}^{\infty} \int_a^b \left(\int_a^b v(t,s)u_j(s)ds \int_a^b v(t,s)u_j(s)ds \right) dt \\
&= \int_a^b \int_a^b v(t,s) \sum_{j=1}^{\infty} \left\langle v(t,\cdot), u_j \right\rangle u_j(s)dsdt \\
&= \int_a^b \int_a^b v^2(t,s)dtds < \infty,
\end{aligned}$$

es decir, $\mathbf{\Gamma}$ es un operador acotado sobre un espacio de Hilbert cuya norma es finita, luego es un operador Hilbert–Schmidt.

Para cualquier variable aleatoria U que definamos como una combinación lineal del proceso $X(t)$, es decir,

$$U = \int_a^b \alpha(t)X(t)dt = \langle \alpha, X\rangle_{L^2[a,b]}, \ \alpha \in L^2[a,b],$$

tenemos que, por el Teorema de Fubini, $Var(U) = \langle \alpha, \mathbf{\Gamma}\alpha\rangle_{L^2[a,b]}$. Veámoslo,

$$\begin{aligned}
Var\,(U) &= E\left(U^2\right) = E\left(\left(\int_a^b \alpha(t)X(t)dt\right)^2\right) = E\left(\int_a^b \alpha(t)X(t)dt \int_a^b \alpha(s)X(s)ds\right) \\
&= E\left(\int_a^b \int_a^b \alpha(t)X(t)X(s)\alpha(s)dsdt\right) = \int_a^b \int_a^b \alpha(t)E(X(t)X(s))\alpha(s)dsdt \\
&= \int_a^b \int_a^b \alpha(t)v(t,s)\alpha(s)dsdt = \langle \alpha, \mathbf{\Gamma}\alpha\rangle_{L^2[a,b]}.
\end{aligned}$$

Se define la primera componente principal como la variable $U_1 = \langle \alpha_1, X \rangle_{L^2[a,b]}$, tal que,

$$Var(U_1) = \sup_{\|\alpha\|_{L^2[a,b]}=1} Var\left(\langle \alpha, X \rangle_{L^2[a,b]}\right) = \sup_{\|\alpha\|_{L^2[a,b]}=1} \langle \alpha, \mathbf{\Gamma}\alpha \rangle_{L^2[a,b]},$$

y la k-ésima componente principal como la variable

$$U_k = \langle \alpha_k, X \rangle_{L^2[a,b]}, \tag{5.8}$$

tal que,

$$\begin{aligned} Var(U_k) &= \sup Var\left(\langle \alpha, X \rangle_{L^2[a,b]}\right) = \sup \langle \alpha, \mathbf{\Gamma}\alpha \rangle_{L^2[a,b]} \\ &\quad \text{sujeto a } \|\alpha\|_{L^2[a,b]} = 1 \text{ y } \left\langle \alpha, \alpha_j \right\rangle_{L^2[a,b]} = 0 \text{ para } j = 1, \ldots, k-1. \end{aligned}$$

Del Teorema de Riesz (Riesz & Nagy (1965)) se concluye que si $\lambda_1 > \lambda_2 > ... > \lambda_j > \lambda_{j+1} > ...$, son los autovalores de $\mathbf{\Gamma}$, entonces las componentes principales se obtienen a partir de sus correspondientes autofunciones. Es decir, sea $\{\alpha_k, k \geq 1\}$ la base de autofunciones del operador lineal de covarianza $\mathbf{\Gamma}$, entonces $U_k = \langle \alpha_k, X \rangle$ es la k-ésima componente principal, y $Var(U_k) = \lambda_k$.

A su vez, notemos que,

$$\|\mathbf{\Gamma}\|^2_{\mathcal{F}} = \langle \mathbf{\Gamma}, \mathbf{\Gamma} \rangle_{\mathcal{F}} = \sum_{k=1}^{\infty} \langle \mathbf{\Gamma}\alpha_k, \mathbf{\Gamma}\alpha_k \rangle_{L^2[a,b]} = \sum_{k=1}^{\infty} \langle \lambda_k\alpha_k, \lambda_k\alpha_k \rangle_{L^2[a,b]} = \sum_{k=1}^{\infty} \lambda_k^2.$$

En lo que sigue asumiremos que todos los autovalores son diferentes.

Definimos para cada $I \in \mathcal{I}_d$, $U_k(I) = \langle \alpha_k, Z(I) \rangle_{L^2[a,b]}$. Asumimos que las primeras $l < p$ componentes principales son suficientes para tener una buena representación del proceso original. El objetivo es encontrar un conjunto I tal que $U_k(I)$ este lo más cerca posible de U_k para todo $k = 1, \ldots, l$.

Definimos la función objetivo del siguiente modo,

$$h(I) = \sum_{k=1}^{l} E\left((U_k - U_k(I))^2\right). \tag{5.9}$$

Esta función mide la suma de las esperanzas de las distancias al cuadrado entre las proyecciones considerando la trayectoria original y las trayectorias *blinded*. Dado $d < p$, buscamos un conjunto $I \in \mathcal{I}_d$ que minimice la función objetivo (5.9), como en (5.2).

Para definir la versión empírica de la función (5.9), necesitamos definir las componentes principales empíricas.

Notamos $v_n(t, s)$ a la función de covarianza empírica, es decir,

$$v_n(t, s) = \frac{1}{n} \sum_{j=1}^{n} X_j(t) X_j(s),$$

y $\mathbf{\Gamma}_n$ a su correspondiente operador lineal, que se define como

$$\mathbf{\Gamma}_n = \frac{1}{n}\sum_{j=1}^{n}\mathbf{V}_j, \tag{5.10}$$

donde $\mathbf{V}_j$ es el operador lineal dado por

$$\left(\mathbf{V}_j u\right)(t) = \int_a^b X_j(t)X_j(s)u(s)ds.$$

Resultando, por el Teorema de Fubini, $E\left(\mathbf{V}_j\right) = \mathbf{\Gamma}$ para todo $1 \leq j \leq n$.

La versión empírica de la función (5.9) está dada por

$$h_n(I) = \sum_{k=1}^{l}\frac{1}{n}\sum_{j=1}^{n}\left(U_k^j - U_k^j(I)\right)^2, \tag{5.11}$$

donde $U_k^j = \left\langle \alpha_k^n, X_j\right\rangle_{L^2[a,b]}$ y $U_k^j(I) = \left\langle \alpha_k^n, \hat{X}_j(I)\right\rangle_{L^2[a,b]}$ y α_k^n son los pesos de la k-ésima componente principal de la muestra aleatoria $\{X_1, \ldots, X_n\}$. Además, por el Teorema de Riesz, $\left\{\alpha_k^n, k \geq 1\right\}$ es la base de autofunciones de $\mathbf{\Gamma}_n$ asociada a $\left\{\lambda_k^n, k \geq 1\right\}$. Elegimos los conjuntos de funciones que minimicen la ecuación (5.11) como en (5.3).

Para obtener el resultado de consistencia, además de necesitar la hipótesis **Hf1**, pedimos las siguientes condiciones adicionales:

Hf2. $E\left(\|X - Z(I)\|^2_{L^2[a,b]}\right) < \infty$.

HfP1. $E\left(\|X\|^2_{L^2[a,b]}\right) < \infty$ y $\|\nu\|_{L^2([a,b]\times[a,b])} = \int_a^b\int_a^b \nu^2(t,s)dtds < \infty$.

Teorema 7 *Sean $\left\{X_j(t) : t \in [a,b]\right\}$ procesos estocásticos independientes e idénticamente distribuidos que satisfacen (5.8). Dado $d, 1 \leq d \leq p$, sea $\mathcal{I}_d$ la familia de todos los subconjuntos de $\{1, \ldots, p\}$ con cardinal d y sea $\mathcal{I}_0 \subset \mathcal{I}_d$ la familia de todos los subconjuntos para los cuales se alcanza el mínimo de la función (5.9). Bajo* **Hf1**, **Hf2** *y* **HfP1** *tenemos que para cada $I_n \in \mathcal{I}_n$, existe un $n_0(\omega)$ tal que para todo $n > n_0(\omega)$, con probabilidad uno, $I_n \in \mathcal{I}_0$.*

La demostración del Teorema 7 se encuentra en el último capítulo.

5.2.3 Modelo Lineal

Como mencionamos en la introducción del capítulo abordamos el problema de regresión lineal cuando la respuesta es escalar y cuando es funcional.

Modelo Lineal con Respuesta Escalar

Sea $Y \in \mathbb{R}$ y $X \in L^2[a,b]$ el modelo lineal con respuesta escalar se define del siguiente modo,

$$Y = \int_a^b \beta(t)X(t)dt + \varepsilon, \tag{5.12}$$

donde $\beta \in L^2[a,b]$ y ε es una variable aleatoria tal que $E(\varepsilon) = 0$ y $E(X(t)\varepsilon) = 0$ para casi todo $t \in [a,b]$.

Consideramos $\beta_0 \in L^2[a,b]$ tal que

$$\beta_0 = arg \min_{\beta \in L^2[a,b]} E\left(\left(Y - \int_a^b \beta(t)X(t)dt\right)^2\right). \tag{5.13}$$

En este contexto, la existencia y unicidad de β_0 no están garantizadas. A continuación damos condiciones bajo las cuales se podrán asegurar estas propiedades.

Sin pérdida de generalidad, asumimos que el proceso estocástico $X(t)$ está centrado, es decir, $E(X(t)) = 0$ para casi todo $t \in [a,b]$ y asumimos que tiene segundo momento finito, es decir, $E\left(\|X\|^2_{L^2[a,b]}\right) < \infty$. Se puede ver por el Teorema de Fubini que $E(Y) = 0$, ya que, $E(X(t)) = 0$ y $E(\varepsilon) = 0$.

Consideramos $\mathbf{\Gamma}$ el operador de covarianza del elemento aleatorio X definido en (5.6). Notemos que $\mathbf{\Gamma}$ verifica que

$$\mathbf{\Gamma}u = E\left(\langle X, u\rangle_{L^2[a,b]} X\right) \text{ para todo } u \in L^2[a,b],$$

ya que,

$$(\mathbf{\Gamma}u)(t) = \int_a^b v(s,t)u(s)ds = E\left(\langle X, u\rangle_{L^2[a,b]} X(t)\right) \quad \text{para todo } t.$$

Utilizando la ecuación (5.12) obtenemos

$$\begin{aligned} E(X(t)Y) &= E\left(X(t)\int_a^b \beta(s)X(s)ds + X(t)\varepsilon\right) \\ &= E\left(X(t)\langle X, \beta\rangle_{L^2[a,b]}\right) + E(X(t)\varepsilon) = (\mathbf{\Gamma}\beta)(t), \end{aligned} \tag{5.14}$$

donde la última igualdad es una consecuencia de $E(X(t)\varepsilon) = 0$ y de la ecuación (5.6).

Llamando $\mathbf{\Delta}$ al operador de covarianza de (Y, X), es decir $\mathbf{\Delta} = E(X(t)Y)$ obtenemos

$$\mathbf{\Delta} = \mathbf{\Gamma}\beta. \tag{5.15}$$

Como mencionamos anteriormente, $\mathbf{\Gamma}$ es un operador Hilbert–Schmidt y no admite inversa continua pues se encuentra en un espacio de dimensión infinita. Por este motivo no se puede hallar en forma explícita una expresión para β.

Además, veamos que la solución de la ecuación (5.15) puede no ser única. El núcleo del operador $\mathbf{\Gamma}$ está dado por

$$Nu(\mathbf{\Gamma}) = \left\{u \in L^2[a,b], \mathbf{\Gamma}u = 0\right\}.$$

Asumimos que $Nu(\mathbf{\Gamma}) \neq \{0\}$ y sean β_0 solución de (5.15) y $\beta_N \in Nu(\mathbf{\Gamma})$. Entonces $\beta_0 + \beta_N$ es también solución de la ecuación (5.15). Por lo tanto, β_0 puede ser único solo en el espacio ortogonal al $Nu(\mathbf{\Gamma})$. Para lidiar con este problema asumiremos que $Nu(\mathbf{\Gamma}) = \{0\}$ o análogamente consideramos β_0 perteneciente a la clausura de $Im(\mathbf{\Gamma}) = \left\{\mathbf{\Gamma}u, u \in L^2[a,b]\right\}$, el cual notaremos $\overline{Im(\mathbf{\Gamma})}$. Veamos que el espacio ortogonal al $Nu(\mathbf{\Gamma})$, al que llamaremos $Nu(\mathbf{\Gamma})^\perp$, es el espacio $\overline{Im(\mathbf{\Gamma})}$,

$$\begin{aligned} u \in Nu(\mathbf{\Gamma}) \text{ sii } \mathbf{\Gamma}u = 0 \text{ sii } \langle \mathbf{\Gamma}u, v\rangle = 0 \text{ para todo } v \in L^2[a,b] \\ \text{sii } \langle u, \mathbf{\Gamma}v\rangle = 0 \text{ para todo } v \in L^2[a,b] \text{ sii } u \in Im(\mathbf{\Gamma})^\perp, \end{aligned}$$

luego $Nu(\mathbf{\Gamma}) = Im(\mathbf{\Gamma})^\perp$, es decir, $Nu(\mathbf{\Gamma})^\perp = (Im(\mathbf{\Gamma})^\perp)^\perp = \overline{Im(\mathbf{\Gamma})}$.

Consideramos $\left\{\alpha_j, j \geq 1\right\}$ la base ortonormal de autofunciones de $\mathbf{\Gamma}$ y $\left\{\lambda_j, j \geq 1\right\}$ los correspondientes autovalores, ordenados en forma decreciente. Por el Teorema de Schmidt, podemos representar a β como $\beta = \sum_{j=1}^{\infty} \left\langle \beta, \alpha_j \right\rangle_{L^2[a,b]} \alpha_j$. Luego, de (5.14) obtenemos,

$$\begin{aligned} \left\langle E(XY), \alpha_j \right\rangle_{L^2[a,b]} &= \left\langle \mathbf{\Gamma}\beta, \alpha_j \right\rangle_{L^2[a,b]} = \left\langle \mathbf{\Gamma} \sum_{k=1}^{\infty} \langle \beta, \alpha_k \rangle_{L^2[a,b]} \alpha_k, \alpha_j \right\rangle_{L^2[a,b]} = \left\langle \sum_{k=1}^{\infty} \langle \beta, \alpha_k \rangle_{L^2[a,b]} \mathbf{\Gamma}\alpha_k, \alpha_j \right\rangle_{L^2[a,b]} \\ &= \sum_{k=1}^{\infty} \left\langle \lambda_k \langle \beta, \alpha_k \rangle_{L^2[a,b]} \alpha_k, \alpha_j \right\rangle_{L^2[a,b]} = \sum_{k=1}^{\infty} \lambda_k \langle \beta, \alpha_k \rangle_{L^2[a,b]} \left\langle \alpha_k, \alpha_j \right\rangle_{L^2[a,b]} \\ &= \lambda_j \left\langle \beta, \alpha_j \right\rangle_{L^2[a,b]}, \quad \text{para } j = 1, 2, \ldots. \end{aligned}$$

Es decir, podemos escribir a β_0 como

$$\beta_0 = \sum_{j=1}^{\infty} \frac{\left\langle E(XY), \alpha_j \right\rangle_{L^2[a,b]}}{\lambda_j} \alpha_j. \tag{5.16}$$

Veamos que $\beta_0 \in L^2[a,b]$. Consideramos la descomposición de Karhunen–Loève para X, es decir,

$$X(t) = \sum_{j=1}^{\infty} \xi_j \alpha_j(t) \text{ para todo } t \in [a,b],$$

donde ξ_j, $j = 1, 2, \ldots$ son variables aleatorias no correlacionadas con esperanza cero y vari-

anza λ_j, y mediante el Teorema de Fubini obtenemos

$$\begin{aligned}\beta_0 &= \sum_{j=1}^{\infty} \frac{\left\langle E\left(\sum_{k=1}^{\infty} \xi_k \alpha_k Y\right), \alpha_j \right\rangle_{L^2[a,b]}}{\lambda_j} \alpha_j = \sum_{j=1}^{\infty} \frac{E\left(\sum_{k=1}^{\infty} \xi_k Y \left\langle \alpha_k, \alpha_j \right\rangle_{L^2[a,b]}\right)}{\lambda_j} \alpha_j \\ &= \sum_{j=1}^{\infty} \frac{E\left(\xi_j Y\right)}{\lambda_j} \alpha_j.\end{aligned}$$

Por lo tanto, $\beta_0 \in L^2[a,b]$ si y solo si X e Y satisfacen

$$\sum_{j=1}^{\infty} \frac{\left(E\left(\xi_j Y\right)\right)^2}{\lambda_j^2} < \infty.$$

Esta condición asegura la existencia y unicidad de la solución, en $\overline{Im\left(\mathbf{\Gamma}\right)}$, del problema (5.13).

Definimos a la función objetivo,

$$h(I) = E\left(\left(\int_a^b \beta_0(t)X(t)dt - \int_a^b \beta_0(t)Z(I)(t)dt\right)^2\right). \tag{5.17}$$

Observemos que esta función mide la esperanza entre la distancia al cuadrado de la función de regresión con el proceso estocástico original y el *blinded*. Dado $d < p$, buscamos un conjunto $I \in \mathcal{I}_d$ que minimice la función objetivo (5.17), como en (5.2).

Para dar la versión empírica del modelo de regresión lineal, es necesario estimar β. Una forma de realizarlo es estimando al operador de covarianza $\mathbf{\Gamma}$. Se proyectan los datos en un espacio de dimensión finita que crezca a medida que aumenta la muestra, el modo más tradicional es utilizando componentes principales. Es decir, se busca estimar β considerando una versión empírica de (5.16) para realizarlo los autoelementos de $\mathbf{\Gamma}$ son reemplazados por los autoelementos del operador de covarianza empírico, teniendo una suma finita. Este método es a veces combinado con algún método de suavizado.

Otro método clásico utilizado se basa en la idea de considerar una base de funciones de $L^2[a,b]$ y se busca un β que verifique la versión empírica de (5.13) adicionándole un término de penalidad para lograr cierta regularidad de la solución. Estas bases no tienen porque ser ortonormales, pueden ser por ejemplo la de Fourier o la de los splines.

Como mencionamos en la Introducción, el problema de como estimar β fue estudiado por diversos autores, entre ellos se encuentran Cardot et al. (2003), Cai & Hall (2006), Hall & Horowitz (2007), Li & Hsing (2007), James et al. (2009). Nosotros consideramos el estimador dado por Cardot et al. (2003), que es un estimador consistente de la función β. Para hallar el estimador hay que seguir una serie de pasos. En primer lugar, se proyectan los datos

en un espacio de dimensión finita generado por los estimadores de las primeras l autofunciones del estimador $\mathbf{\Gamma}$. Es decir, consideramos el espacio generado por $\left\{\alpha_1^n, \ldots, \alpha_l^n\right\}$ donde α_k^n es la autofunción asociada con el k-ésimo mayor autovalor λ_k^n, con $k = 1, \ldots, l$.

Recordemos que definimos la versión empírica de $\mathbf{\Gamma}$ en (5.10), luego

$$\mathbf{\Gamma}_n u(t) = \frac{1}{n}\sum_{j=1}^{n} \mathbf{V}_j u(t) = \frac{1}{n}\sum_{j=1}^{n} \int_a^b X_j(t)X_j(s)u(s)ds = \frac{1}{n}\sum_{j=1}^{n} \left\langle X_j, u\right\rangle_{L^2[a,b]} X_j(t).$$

Del mismo modo, podemos definir la versión empírica de $\mathbf{\Delta}$,

$$\mathbf{\Delta}_n u = \frac{1}{n}\sum_{j=1}^{n} D_j u = \frac{1}{n}\sum_{j=1}^{n} \int_a^b Y_j X_j(s)u(s)ds = \frac{1}{n}\sum_{j=1}^{n} \left\langle X_j, u\right\rangle_{L^2[a,b]} Y_j.$$

Cardot et al. (1999) definen el estimador de β_0 como

$$\beta_{PCR} = \sum_{j=1}^{l} \frac{\mathbf{\Delta}_n \alpha_j^n}{\lambda_j^n} \alpha_j^n,$$

y prueban resultados de convergencia en probabilidad y casi segura. Mediante simulaciones se puede ver que aún para n muy grande β_{PCR} es muy irregular. Con el objetivo de solucionar este problema, es decir, suavizar la curva dada por β_{PCR}, en el año 2003, Cardot et al. (2003) proponen regularizar la curva mediante B-spline. Más específicamente, sean q, k enteros, sea S_{qk} el espacio de las funciones de splines definidas en el intervalo $[a, b]$, de grado q y con $k - 1$ nodos equiespaciados, luego S_{qk} tiene dimensión $q + k$. El estimador de β_0 es

$$\beta_{SPCR} = arg \min_{\beta \in S_{qk}} \int_a^b \left(\beta_{PCR}(t) - \beta_0(t)\right)^2 dt.$$

En su trabajo demuestran, bajo hipótesis de regularidad, la convergencia en probabilidad y casi segura de β_{SPCR} a β_0.

Luego, la versión empírica de la función objetivo (5.17) está dada por

$$h_n(I) = \frac{1}{n}\sum_{j=1}^{n} \left(\int_a^b \beta_n(t)X_j(t)dt - \int_a^b \beta_n(t)\hat{X}_j(I)(t)dt \right)^2, \tag{5.18}$$

donde β_n es el estimador de β_0. Elegimos los conjuntos de funciones que minimicen la ecuación (5.18) como en (5.3).

Para establecer el resultado de consistencia además de necesitar las hipótesis **Hf1** y **Hf2**, requerimos los siguientes supuestos:

HfR1. $\|\beta_n - \beta_0\|_{L^2[a,b]} \to_{c.s.} 0$.

HfR2. $E\left(\|X\|^2_{L^2[a,b]}\right) < \infty$.

Observemos que el estimador β_n propuesto por Cardot et al. (2003) cumple **HfR1**.

Luego tenemos el siguiente resultado de consistencia.

Teorema 8 *Sean* $\left\{\left(Y_j, X_j(t)\right) \in \mathbb{R} \times L^2[a,b]\right\}$ *procesos estocásticos independientes e idénticamente distribuidos que satisfacen (5.12). Dado* $d, 1 \leq d \leq p$, *sea* $\mathcal{I}_d$ *la familia de todos los subconjuntos de* $\{1, \ldots, p\}$ *con cardinal d y sea* $\mathcal{I}_0 \subset \mathcal{I}_d$ *la familia de los conjuntos para los cuales se alcanza el mínimo de la función objetivo (5.17). Bajo* **Hf1**, **Hf2**, **HfR1** *y* **HfR2** *tenemos que para cada* $I_n \in \mathcal{I}_n$, *existe un* $n_0(\omega)$ *tal que para todo* $n > n_0(\omega)$, *con probabilidad uno,* $I_n \in \mathcal{I}_0$.

La demostración del Teorema 8 se encuentra en el último capítulo.

Modelo Lineal con Respuesta Funcional

Sea $Y \in L^2[c,d]$ y $X \in L^2[a,b]$ el modelo de regresión lineal con respuesta funcional, se define del siguiente modo,

$$Y(s) = \int_a^b \beta(t,s)X(t)dt + \varepsilon(s) \quad s \in [c,d], \tag{5.19}$$

donde $\beta \in L^2([a,b] \times [c,d])$ y $\varepsilon(s)$ es una variable aleatoria para cada s tal que $E(\varepsilon(s)) = 0$ y $E(X(t)\varepsilon(s)) = 0$ para casi todo $t \in [a,b]$, $s \in [c,d]$.

Consideramos $\beta_0 \in L^2([a,b] \times [c,d])$ tal que

$$\beta_0 = arg \min_{\beta \in L^2([a,b]\times[c,d])} E\left(\left\|Y - \int_a^b \beta(t,\cdot)X(t)dt\right\|^2_{L^2[c,d]}\right).$$

En este contexto, la existencia y unicidad de β_0 no están garantizadas. A continuación veremos que siguiendo las mismas idea que para el caso de respuesta escalar podremos asegurar la existencia y unicidad de la solución en el espacio del núcleo del operador de covarianza de X.

Sin pérdida de generalidad, asumimos de que $E(X(t)) = 0$ para casi todo $t \in [a,b]$, luego $E(Y(s)) = 0$ para casi todo $s \in [c,d]$. Asumimos $E\left(\|X\|^2_{L^2[a,b]}\right) < \infty$ y $E\left(\|Y\|^2_{L^2[c,d]}\right) < \infty$.

Consideramos $\mathbf{\Gamma}_X$ el operador de covarianza lineal del proceso X y $\mathbf{\Gamma}_Y$ el operador de covarianza lineal del proceso Y. Sea $\left\{\alpha_{X,j}, j \geq 1\right\}$ la base ortonormal de autofunciones de $\mathbf{\Gamma}_X$ y $\left\{\lambda_{X,j}, j \geq 1\right\}$ los correspondientes autovalores, ordenados en forma decreciente. Análogamente consideramos $\left\{\alpha_{Y,j'}, j' \geq 1\right\}$ la base ortonormal de autofunciones de $\mathbf{\Gamma}_Y$ y $\left\{\lambda_{Y,j'}, j' \geq 1\right\}$ los correspondientes autovalores ordenados en forma decreciente.

Repitiendo el mismo análisis que para el caso de respuesta escalar obtenemos que,

$$\beta_0 = \sum_{j,j'=1}^{\infty} \frac{\left\langle E\left(XY\right), \alpha_{X,j}\alpha_{Y,j'}\right\rangle_{L^2([a,b]\times[c,d])}}{\lambda_{X,j}} \alpha_{X,j}\alpha_{Y,j'}. \tag{5.20}$$

Para ver que $\beta_0 \in L^2([a,b]\times[c,d])$ escribimos al proceso X y al proceso Y mediante la descomposición de Karhunen–Loève, es decir,

$$X(t) = \sum_{j=1}^{\infty} \xi_{X,j}\alpha_{X,j}(t) \text{ para todo } t \in [a,b]$$

y

$$Y(s) = \sum_{j'=1}^{\infty} \xi_{Y,j'}\alpha_{Y,j'}(s) \text{ para todo } s \in [c,d].$$

Por lo tanto, $\beta_0 \in L^2([a,b]\times[c,d])$ si y solo si X e Y satisfacen,

$$\sum_{j,j'=1}^{\infty} \frac{\left(E\left(\xi_{X,j}\xi_{Y,j'}\right)\right)^2}{\lambda_{X,j}^2} < \infty.$$

Esta condición asegura la existencia y unicidad de la solución en el espacio ortogonal del núcleo del operador de covarianza de X, $\mathbf{\Gamma}_X$, que al ser el operador autoadjunto coincide con la clausura de la imagen del operador.

En este caso, siguiendo la misma idea que en el modelo lineal con respuesta escalar, la función objetivo es

$$h(I) = E\left(\left\|\int_a^b \beta_0(t,s)X(t)dt - \int_a^b \beta_0(t,s)Z(I)(t)dt\right\|^2_{L^2[c,d]}\right). \tag{5.21}$$

Observemos que al igual que en el caso de respuesta escalar mide la esperanza entre la distancia al cuadrado entre la función de regresión con el proceso estocástico original y el *blinded*. Luego, dado $d < p$, buscamos un conjunto $I \in \mathcal{I}_d$ que minimice la función objetivo (5.21), como indicamos en (5.2).

Para definir la versión empírica, necesitamos estimar a β, este problema fue estudiado por diversos autores, entre ellos, Yao et al. (2005) y Müller & Yao (2008). Yao et al. (2005), proponen el siguiente estimador

$$\beta_n(t,s) = \sum_{j'=1}^{J'} \sum_{j=1}^{J} \frac{\sigma^n_{jj'}\alpha^n_{X,j}(t)\alpha^n_{Y,j'}(s)}{\lambda^n_{X,j}},$$

donde $\sigma^n_{jj'}$ es un estimador de $E\left(\xi_{X,j}\xi_{Y,j'}\right)$. El estimador más natural es

$$\sigma^n_{jj'} = \frac{1}{N}\sum_{i=1}^{N}\left\langle X_i, \alpha^n_{Y,j'}\right\rangle_{L^2([a,b]\times[c,d])}\left\langle Y_i, \alpha^n_{X,j}\right\rangle_{L^2([a,b]\times[c,d])}.$$

En su trabajo demuestran la convergencia en probabilidad de β_n a β_0.

Por otro lado, la versión empírica de la función (5.21) está dada por

$$h_n(I) = \frac{1}{n}\sum_{j=1}^{n}\left\|\int_a^b \beta_n(t,s)X_j(t)dt - \int_a^b \beta_n(t,s)\hat{X}_j(I)(t)dt\right\|^2_{L^2[c,d]}. \tag{5.22}$$

Luego, elegimos los conjuntos de funciones que minimicen la función (5.22) como indicamos en (5.3).

Para establecer el resultado de consistencia, además de requerir las hipótesis **Hf1** y **Hf2** ya establecidas, precisamos las siguientes condiciones adicionales:

HfRf1. $\|\beta_n - \beta_0\|_{L^2([a,b]\times[c,d])} \to_{p.} 0$.

HfRf2. $E\left(\|X\|^2_{L^2[a,b]}\right) < \infty$ y $E\left(\|Y\|^2_{L^2[c,d]}\right) < \infty$.

Notemos que el estimador β_n propuesto por Yao et al. (2005) satisface **HfRf1**.

Luego tenemos el siguiente resultado de consistencia.

Teorema 9 *Sean* $\left\{\left(Y_j(s), X_j(t)\right) \in L^2[c,d]\times L^2[a,b]\right\}$ *procesos estocásticos independientes e idénticamente distribuidos que satisfacen (5.19). Dado* $d, 1 \le d \le p$, *sea* $\mathcal{I}_d$ *la familia de todos los subconjuntos de* $\{1,\dots,p\}$ *con cardinal d y sea* $\mathcal{I}_0 \subset \mathcal{I}_d$ *la familia de todos los subconjuntos para los cuales se alcanza el mínimo de la función (5.21). Bajo* **Hf1**, **Hf2**, **HfRf1** *y* **HfRf2** *tenemos que dado* $I_n \in \mathcal{I}_n$ *se verifica que* $P(I_n \in \mathcal{I}_0) \to 1$.

La demostración del Teorema 9 se encuentra en el último capítulo.

Observación 10 *Si tenemos un estimador* β_n *de* β_0 *que es fuertemente consistente, pudiendo intercambiar la hipótesis* **HfRf1** *por*

HfRf1*. $\|\beta_n - \beta_0\|_{L^2([a,b]\times[c,d])} \to_{c.s.} 0$.

vale el Teorema 9 cambiando la conclusión por: Bajo **Hf1**, **Hf2**, **HfRf1*** *y* **HfRf2** *tenemos que dado* $I_n \in \mathcal{I}_n$, *existe un* $n_0(\omega)$ *tal que para todo* $n > n_0(\omega)$, *con probabilidad uno,* $I_n \in \mathcal{I}_0$.

Capítulo 6

Demostraciones de los Resultados

6.1 Demostraciones Correspondientes al Capítulo 3

6.1.1 Demostración del Teorema 1

Para demostrar el Teorema 1 consideremos previamente el siguiente Lema:

Lema 1 *Si*

$$h_n(I) \rightarrow_{c.s.} h(I) \text{ para todo } I \in \mathcal{I}_d, \tag{6.1}$$

entonces

$$arg \min_{I \in \mathcal{I}_d} h_n(I) \rightarrow_{c.s.} arg \min_{I \in \mathcal{I}_d} h(I).$$

Demostración del Lema 1 En primer lugar observemos que hay un número finito de conjuntos pertenecientes a $\mathcal{I}_d$, es decir la convergencia (6.1) es uniforme. En consecuencia para todo $\delta > 0$ existe $n_0(\omega)$ tal que con probabilidad uno si $n \geq n_0(\omega)$

$$|h_n(I) - h(I)| < \delta \text{ para todo } I \in \mathcal{I}_d,$$

es decir,

$$h(I) - \delta < h_n(I) < h(I) + \delta \text{ para todo } I \in \mathcal{I}_d \tag{6.2}$$

y

$$h_n(I) - \delta < h(I) < h_n(I) + \delta \text{ para todo } I \in \mathcal{I}_d. \tag{6.3}$$

Por un lado, de (3.3), sabemos que existe $\delta_0 > 0$ tal que

$$h(I_0) < h(I) - \delta_0 \text{ para todo } I \notin \mathcal{I}_0, \text{ para todo } I_0 \in \mathcal{I}_0.$$

Sea $\delta = \frac{\delta_0}{2}$, luego de (6.2) obtenemos

$$h_n(I_0) < h(I_0) + \frac{\delta_0}{2} < h(I) - \delta_0 + \frac{\delta_0}{2} < h_n(I) \text{ para todo } I \notin \mathcal{I}_0, \text{ para todo } I_0 \in \mathcal{I}_0, \text{ si } n \geq n_0(\omega)$$

con probabilidad uno. Es decir, existe $n_0(\omega)$ tal que si $n \geq n_0(\omega)$ con probabilidad uno si elegimos $I_0 \in \mathcal{I}_0$, I_0 minimiza $h_n(I)$.

Por otro lado, de (3.7), tenemos que para cada n fijo existe $\delta_1 > 0$ tal que

$$h_n(I_n) < h_n(I) - \delta_1 \text{ para todo } I \notin \mathcal{I}_n, \text{ para todo } I_n \in \mathcal{I}_n.$$

Si consideramos $\delta = \frac{\delta_1}{2}$ en (6.3) y fijamos un $n \geq n_0(\omega)$ obtenemos

$$h(I_n) < h_n(I_n) + \frac{\delta_1}{2} < h_n(I) - \delta_1 + \frac{\delta_1}{2} < h(I) \text{ para todo } I \notin \mathcal{I}_n, \text{ para todo } I_n \in \mathcal{I}_n$$

con probabilidad uno. Es decir, existe $n_0(\omega)$ tal que si $n \geq n_0(\omega)$ para todo $I_n \in \mathcal{I}_n$, I_n minimiza $h(I)$ con probabilidad uno.

Luego, queda demostrado el Lema 1.

■

Por lo tanto, para demostrar el Teorema 1 es suficiente ver que para cada conjunto fijo I la función objetivo empírica (3.6) converge c.s. a la función objetivo teórica (3.2). Para probarlo veamos que

$$\left\|\alpha_k^n(P_n) - \alpha_k^n(Q_n(I))\right\| \to_{c.s.} \|\alpha_k(P) - \alpha_k(Q(I))\| \text{ para todo } k.$$

Dauxois et al. (1982) demostraron que bajo **HP1**, si

$$\sup_{\|u\|=1} \left\|\left(\hat{\Sigma}_n - \Sigma\right)(u)\right\| \to_{c.s.} 0,$$

entonces

$$\left\|\alpha_k^n(P_n) - \alpha_k(P)\right\| \to_{c.s.} 0 \text{ para todo } 1 \leq k \leq p,$$

donde $\alpha_k^n(P_n)$ (respec. $\alpha_k(P)$) son los autovectores de $\hat{\Sigma}_n$, que es la matriz de covarianza empírica asociada a P_n (respec. Σ es la matriz de covarianza asociada a P).

Luego, si mostramos que

$$\sup_{\|u\|=1} \left\|\left(\hat{\Sigma}_n(I) - \Sigma(I)\right)(u)\right\| \to_{c.s.} 0,$$

entonces

$$\left\|\alpha_k^n(Q_n(I)) - \alpha_k(Q(I))\right\| \to_{c.s.} 0 \text{ para todo } 1 \leq k \leq p,$$

donde $\alpha_k^n(Q_n(I))$ (respec. $\alpha_k(Q(I))$) son los autovectores de $\hat{\Sigma}_n(I)$, que es la matriz de covarianza empírica asociada a $Q_n(I)$ (respec. $\Sigma(I)$ es la matriz de covarianza asociada a $Q(I)$).

Observamos que el caso $I = \{1, \ldots, p\}$ es el tradicional, pero queremos ver que para cualquier subconjunto $I \subseteq \{1, \ldots, p\}$ vale que

$$\sup_{\|u\|=1} \left\| \left(\hat{\Sigma}_n(I) - \Sigma(I) \right)(u) \right\| \to_{c.s.} 0. \tag{6.4}$$

Para demostrarlo definimos los vectores aleatorios no observables, $\mathbf{Z}_1(I), \ldots, \mathbf{Z}_n(I)$, donde

$$Z_j(I)[i] = \begin{cases} X_j[i] & \text{si } i \in I, \\ E\left(X_j[i] | \mathbf{X}_j(I)\right) & \text{si } i \notin I, \end{cases} \tag{6.5}$$

y notamos $Q_n^*(I)$ su distribución empírica y $\Sigma_n^*(I)$ su matriz de covarianza. Probaremos que

$$\sup_{\|u\|=1} \left\| \left(\hat{\Sigma}_n(I) - \Sigma_n^*(I) \right)(u) \right\| \to_{c.s.} 0 \tag{6.6}$$

y

$$\sup_{\|u\|=1} \left\| \left(\Sigma_n^*(I) - \Sigma(I) \right)(u) \right\| \to_{c.s.} 0. \tag{6.7}$$

Para demostrar (6.6) y (6.7), al estar en un espacio de dimensión finita, alcanza con ver que las diferencias entre las matrices de covarianza convergen a cero en cada una de sus coordenadas.

Para simplificar la notación, podemos asumir, sin pérdida de generalidad, que $I = \{1, \ldots, d\}$, entonces,

$$\hat{X}_j(I)[i] = \begin{cases} X_j[i] & \text{si } 1 \leq i \leq d, \\ \eta_n^i\left(\mathbf{X}_j(I)\right) & \text{si } d+1 \leq i \leq p, \end{cases}$$

donde $\eta_n^i(\mathbf{X}_j(I))$ verifica **H1**, es decir, es un estimador no paramétrico fuertemente consistente de la esperanza condicional, específicamente,

$$\eta_n^i\left(\mathbf{X}_j(I)\right) \to_{c.s.} E\left(X_j[i] | \mathbf{X}_j(I)\right) \text{ para todo } 1 \leq j \leq n,\ d+1 \leq i \leq p. \tag{6.8}$$

Cada coordenada de las matrices de covarianza está dada por,

$$\left(\hat{\Sigma}_n(I)\right)_{ii'} = \begin{cases} \frac{1}{n}\sum_{j=1}^n \left(X_j[i] - \overline{X[i]}\right)\left(X_j[i'] - \overline{X[i']}\right) & \text{si } 1 \leq i, i' \leq d, \\ \frac{1}{n}\sum_{j=1}^n \left(X_j[i] - \overline{X[i]}\right)\left(\eta_n^{i'}\left(\mathbf{X}_j(I)\right) - \overline{\eta_n^{i'}\left(\mathbf{X}(I)\right)}\right) & \text{si } 1 \leq i \leq d < i' \leq p, \\ \frac{1}{n}\sum_{j=1}^n \left(\eta_n^{i}\left(\mathbf{X}_j(I)\right) - \overline{\eta_n^{i}\left(\mathbf{X}(I)\right)}\right)\left(X_j[i'] - \overline{X[i']}\right) & \text{si } 1 \leq i' \leq d < i \leq p, \\ \frac{1}{n}\sum_{j=1}^n \left(\eta_n^{i}\left(\mathbf{X}_j(I)\right) - \overline{\eta_n^{i}\left(\mathbf{X}(I)\right)}\right)\left(\eta_n^{i'}\left(\mathbf{X}_j(I)\right) - \overline{\eta_n^{i'}\left(\mathbf{X}(I)\right)}\right) & \text{si } d+1 \leq i, i' \leq p, \end{cases}$$

$$(\Sigma_n^*(I))_{ii'} = \begin{cases} \frac{1}{n}\sum_{j=1}^n \left(X_j[i] - \overline{X[i]}\right)\left(X_j[i'] - \overline{X[i']}\right) \text{ si } 1 \le i, i' \le d, \\ \frac{1}{n}\sum_{j=1}^n \left(X_j[i] - \overline{X[i]}\right)\left(E\left(X_j[i']|\mathbf{X}_j(I)\right) - \overline{E\left(X[i']|\mathbf{X}(I)\right)}\right) \text{ si } 1 \le i \le d < i' \le p, \\ \frac{1}{n}\sum_{j=1}^n \left(E\left(X_j[i]|\mathbf{X}_j(I)\right) - \overline{E\left(X[i]|\mathbf{X}(I)\right)}\right)\left(X_j[i'] - \overline{X[i']}\right) \text{ si } 1 \le i' \le d < i \le p, \\ \frac{1}{n}\sum_{j=1}^n \left(E\left(X_j[i]|\mathbf{X}_j(I)\right) - \overline{E\left(X[i]|\mathbf{X}(I)\right)}\right)\left(E\left(X_j[i']|\mathbf{X}_j(I)\right) - \overline{E\left(X[i']|\mathbf{X}(I)\right)}\right) \\ \qquad\qquad \text{si } d+1 \le i, i' \le p, \end{cases}$$

y

$$(\Sigma(I))_{ii'} = \begin{cases} cov\left(X[i], X[i']\right) \text{ si } 1 \le i, i' \le d, \\ cov\left(X[i], E(X[i']|\mathbf{X}(I))\right) \text{ si } 1 \le i \le d < i' \le p, \\ cov\left(E(X[i]|\mathbf{X}(I)), X[i']\right) \text{ si } 1 \le i' \le d < i \le p, \\ cov\left(E(X[i]|\mathbf{X}(I)), E(X[i']|\mathbf{X}(I))\right) \text{ si } d+1 \le i, i' \le p, \end{cases}$$

donde

$$\overline{X[i]} = \frac{1}{n}\sum_{j=1}^n X_j[i], \qquad \overline{\eta_n^i\left(\mathbf{X}(I)\right)} = \frac{1}{n}\sum_{j=1}^n \eta_n^i\left(\mathbf{X}_j(I)\right),$$

$$\overline{E\left(X[i]|\mathbf{X}(I)\right)} = \frac{1}{n}\sum_{j=1}^n E\left(X_j[i]|\mathbf{X}_j(I)\right).$$

Primero demostremos (6.6), viendo que las diferencias de las coordenadas de las matrices $\hat{\Sigma}_n(I)$ y $\Sigma_n^*(I)$ convergen a cero. Para probarlo consideramos cuatro casos diferentes para las coordenadas de $\left(\hat{\Sigma}_n(I) - \Sigma_n^*(I)\right)_{ii'}$.

- Si $1 \le i, i' \le d$

$$\frac{1}{n}\sum_{j=1}^n \left(X_j[i] - \overline{X[i]}\right)\left(X_j[i'] - \overline{X[i']}\right) - \frac{1}{n}\sum_{j=1}^n \left(X_j[i] - \overline{X[i]}\right)\left(X_j[i'] - \overline{X[i']}\right) = 0.$$

- Si $1 \le i \le d < i' \le p$

$$\begin{aligned} & \frac{1}{n}\sum_{j=1}^n \left(X_j[i] - \overline{X[i]}\right)\left(\eta_n^{i'}\left(\mathbf{X}_j(I)\right) - \overline{\eta_n^{i'}\left(\mathbf{X}(I)\right)}\right) - \\ & \frac{1}{n}\sum_{j=1}^n \left(X_j[i] - \overline{X[i]}\right)\left(E\left(X_j[i']|\mathbf{X}_j(I)\right) - \overline{E\left(X[i']|\mathbf{X}(I)\right)}\right) \\ = \; & \frac{1}{n}\sum_{j=1}^n \left(X_j[i] - \overline{X[i]}\right)\left(\eta_n^{i'}\left(\mathbf{X}_j(I)\right) - E\left(X_j[i']|\mathbf{X}_j(I)\right) - \overline{\eta_n^{i'}\left(\mathbf{X}(I)\right)} + \overline{E\left(X[i']|\mathbf{X}(I)\right)}\right). \end{aligned}$$

De la desigualdad de Cauchy–Schwarz tenemos que,

$$\left|\frac{1}{n}\sum_{j=1}^{n}\left(X_j[i]-\overline{X[i]}\right)\left(\eta_n^{i'}\left(\mathbf{X}_j(I)\right)-E\left(X_j[i']|\mathbf{X}_j(I)\right)-\overline{\eta_n^{i'}\left(\mathbf{X}(I)\right)}+\overline{E\left(X[i']|\mathbf{X}(I)\right)}\right)\right|$$

$$\leq \left(\frac{1}{n}\sum_{j=1}^{n}\left(X_j[i]-\overline{X[i]}\right)^2\right)^{\frac{1}{2}}\left(\frac{1}{n}\sum_{j=1}^{n}\left(\eta_n^{i'}\left(\mathbf{X}_j(I)\right)-E\left(X_j[i']|\mathbf{X}_j(I)\right)-\overline{\eta_n^{i'}\left(\mathbf{X}(I)\right)}+\overline{E\left(X[i']|\mathbf{X}(I)\right)}\right)^2\right)^{\frac{1}{2}}.$$

Por un lado, por la Ley Fuerte de los Grandes Números

$$\frac{1}{n}\sum_{j=1}^{n}\left(X_j[i]-\overline{X[i]}\right)^2 \to_{c.s.} Var(X[i]),$$

y $Var(X[i])$ es finita por **HP1**, ya que $E\left(||X||^2\right)$ es finita. Por el otro, por (6.8),

$$\frac{1}{n}\sum_{j=1}^{n}\left(\eta_n^{i'}\left(\mathbf{X}_j(I)\right)-E\left(X_j[i']|\mathbf{X}_j(I)\right)-\overline{\eta_n^{i'}\left(\mathbf{X}(I)\right)}+\overline{E\left(X[i']|\mathbf{X}(I)\right)}\right)^2 \to_{c.s.} 0,$$

por consiguiente

$$\left|\frac{1}{n}\sum_{j=1}^{n}\left(X_j[i]-\overline{X[i]}\right)\left(\eta_n^{i'}\left(\mathbf{X}_j(I)\right)-E\left(X_j[i']|\mathbf{X}_j(I)\right)-\overline{\eta_n^{i'}\left(\mathbf{X}(I)\right)}+\overline{E\left(X[i']|\mathbf{X}(I)\right)}\right)\right| \to_{c.s.} 0.$$

- Si $1 \leq i' \leq d < i \leq p$

$$\frac{1}{n}\sum_{j=1}^{n}\left(\eta_n^{i}\left(\mathbf{X}_j(I)\right)-\overline{\eta_n^{i}\left(\mathbf{X}(I)\right)}\right)\left(X_j[i']-\overline{X[i']}\right)-$$

$$\frac{1}{n}\sum_{j=1}^{n}\left(E\left(X_j[i]|\mathbf{X}_j(I)\right)-\overline{E\left(X[i]|\mathbf{X}(I)\right)}\right)\left(X_j[i']-\overline{X[i']}\right)$$

$$= \frac{1}{n}\sum_{j=1}^{n}\left(\eta_n^{i}\left(\mathbf{X}_j(I)\right)-E\left(X_j[i]|\mathbf{X}_j(I)\right)-\overline{\eta_n^{i}\left(\mathbf{X}(I)\right)}+\overline{E\left(X[i]|\mathbf{X}(I)\right)}\right)\left(X_j[i']-\overline{X[i']}\right),$$

que converge a cero c.s. La demostración es análoga al item anterior.

- Si $d+1 \leq i, i' \leq p$

$$\frac{1}{n}\sum_{j=1}^{n}\left(\eta_n^{i}\left(\mathbf{X}_j(I)\right)-\overline{\eta_n^{i}\left(\mathbf{X}(I)\right)}\right)\left(\eta_n^{i'}\left(\mathbf{X}_j(I)\right)-\overline{\eta_n^{i'}\left(\mathbf{X}(I)\right)}\right)-$$

$$\frac{1}{n}\sum_{j=1}^{n}\left(E\left(X_j[i]|\mathbf{X}_j(I)\right)-\overline{E\left(X[i]|\mathbf{X}(I)\right)}\right)\left(E\left(X_j[i']|\mathbf{X}_j(I)\right)-\overline{E\left(X[i']|\mathbf{X}(I)\right)}\right)$$

$$= \frac{1}{n}\sum_{j=1}^{n}\Big[\left(\eta_n^{i}\left(\mathbf{X}_j(I)\right)-\overline{\eta_n^{i}\left(\mathbf{X}(I)\right)}\right)\left(\eta_n^{i'}\left(\mathbf{X}_j(I)\right)-\overline{\eta_n^{i'}\left(\mathbf{X}(I)\right)}\right)-$$

$$\left(E\left(X_j[i]|\mathbf{X}_j(I)\right)-\overline{E\left(X[i]|\mathbf{X}(I)\right)}\right)\left(E\left(X_j[i']|\mathbf{X}_j(I)\right)-\overline{E\left(X[i']|\mathbf{X}(I)\right)}\right)\Big],$$

que converge a cero c.s., ya que por (6.8) tenemos que

$$\eta_n^i\left(\mathbf{X}_j(I)\right) - \overline{\eta_n^i\left(\mathbf{X}(I)\right)} \to_{c.s.} E\left(X_j[i]|\mathbf{X}_j(I)\right) - \overline{E\left(X[i]|\mathbf{X}(I)\right)}$$

y

$$\eta_n^{i'}\left(\mathbf{X}_j(I)\right) - \overline{\eta_n^{i'}\left(\mathbf{X}(I)\right)} \to_{c.s.} E\left(X_j[i']|\mathbf{X}_j(I)\right) - \overline{E\left(X[i']|\mathbf{X}(I)\right)}.$$

En consecuencia, mostramos que todas las coordenadas de la matriz $\hat{\Sigma}_n(I) - \Sigma_n^*(I)$ convergen a cero c.s., por lo tanto probamos (6.6).

En segundo lugar veamos (6.7), para demostrarlo probaremos, al igual que antes, que cada una de las coordenadas de la matriz $\Sigma_n^*(I) - \Sigma(I)$ converge a cero. Nuevamente consideramos los cuatro casos para $\left(\Sigma_n^*(I) - \Sigma(I)\right)_{ii'}$.

- Si $1 \leq i, i' \leq d$

$$\frac{1}{n}\sum_{j=1}^{n}\left(X_j[i] - \overline{X[i]}\right)\left(X_j[i'] - \overline{X[i']}\right) - cov\left(X[i], X[i']\right) \to_{c.s.} 0.$$

- Si $1 \leq i \leq d < i' \leq p$

$$\frac{1}{n}\sum_{j=1}^{n}\left(X_j[i] - \overline{X[i]}\right)\left(E\left(X_j[i']|\mathbf{X}_j(I)\right) - \overline{E\left(X[i']|\mathbf{X}(I)\right)}\right) - cov\left(X[i], E\left(X[i']|\mathbf{X}(I)\right)\right) \to_{c.s.} 0.$$

- Si $1 \leq i' \leq d < i \leq p$

$$\frac{1}{n}\sum_{j=1}^{n}\left(E\left(X_j[i]|\mathbf{X}_j(I)\right) - \overline{E\left(X[i]|\mathbf{X}(I)\right)}\right)\left(X_j[i'] - \overline{X[i']}\right) - cov\left(E\left(X[i]|\mathbf{X}(I)\right), X[i']\right) \to_{c.s.} 0.$$

- Si $d+1 \leq i, i' \leq p$

$$\frac{1}{n}\sum_{j=1}^{n}\left(E\left(X_j[i]|\mathbf{X}_j(I)\right) - \overline{E\left(X[i]|\mathbf{X}(I)\right)}\right)\left(E\left(X_j[i']|\mathbf{X}_j(I)\right) - \overline{E\left(X[i']|\mathbf{X}(I)\right)}\right) -$$
$$cov\left(E\left(X[i]|\mathbf{X}(I)\right), E\left(X[i']|\mathbf{X}(I)\right)\right) \to_{c.s.} 0.$$

Por consiguiente demostramos que todas las coordenadas de la matriz $\Sigma_n^*(I) - \Sigma(I)$ convergen a cero, por lo tanto (6.7) quedó demostrado.

Luego (6.4) queda demostrado como consecuencia de (6.6) y (6.7). Por Dauxois et al. (1982) tenemos que

$$\left\|\alpha_k^n(Q_n(I)) - \alpha_k(Q(I))\right\| \to_{c.s.} 0 \text{ para todo } 1 \leq k \leq p$$

y

$$\left\|\alpha_k^n(P_n) - \alpha_k(P)\right\| \to_{c.s.} 0 \text{ para todo } 1 \leq k \leq p.$$

Por la desigualdad triangular, sabemos que

$$\begin{aligned}
&\left\|\alpha_k^n(P_n) - \alpha_k^n(Q_n(I))\right\| - \|\alpha_k(P) - \alpha_k(Q(I))\| \\
= &\left\|\alpha_k^n(P_n) - \alpha_k(P) + \alpha_k(P) - \alpha_k(Q(I)) + \alpha_k(Q(I)) - \alpha_k^n(Q_n(I))\right\| - \|\alpha_k(P) - \alpha_k(Q(I))\| \\
\leq &\left\|\alpha_k^n(P_n) - \alpha_k(P)\right\| + \|\alpha_k(P) - \alpha_k(Q(I))\| + \left\|\alpha_k(Q(I)) - \alpha_k^n(Q_n(I))\right\| - \|\alpha_k(P) - \alpha_k(Q(I))\| \\
= &\left\|\alpha_k^n(P_n) - \alpha_k(P)\right\| + \left\|\alpha_k^n(Q_n(I)) - \alpha_k(Q(I))\right\|
\end{aligned}$$

y

$$\begin{aligned}
&\|\alpha_k(P) - \alpha_k(Q(I))\| - \left\|\alpha_k^n(P_n) - \alpha_k^n(Q_n(I))\right\| \\
= &\left\|\alpha_k(P) - \alpha_k^n(P_n) + \alpha_k^n(P_n) - \alpha_k^n(Q_n(I)) + \alpha_k^n(Q_n(I)) - \alpha_k(Q(I))\right\| - \left\|\alpha_k^n(P_n) - \alpha_k^n(Q_n(I))\right\| \\
\leq &\left\|\alpha_k(P) - \alpha_k^n(P_n)\right\| + \left\|\alpha_k^n(P_n) - \alpha_k^n(Q_n(I))\right\| + \left\|\alpha_k^n(Q_n(I)) - \alpha_k(Q(I))\right\| - \left\|\alpha_k^n(P_n) - \alpha_k^n(Q_n(I))\right\| \\
= &\left\|\alpha_k^n(P_n) - \alpha_k(P)\right\| + \left\|\alpha_k^n(Q_n(I)) - \alpha_k(Q(I))\right\|
\end{aligned}$$

Luego,

$$\begin{aligned}
&\left|\left\|\alpha_k^n(P_n) - \alpha_k^n(Q_n(I))\right\| - \|\alpha_k(P) - \alpha_k(Q(I))\|\right| \\
\leq &\left\|\alpha_k^n(P_n) - \alpha_k(P)\right\| + \left\|\alpha_k^n(Q_n(I)) - \alpha_k(Q(I))\right\|,
\end{aligned}$$

donde cada una de los términos converge a cero c.s. y de esta forma quedó demostrado que $h_n(I) \to_{c.s.} h(I)$.

6.2 Demostraciones Correspondientes al Capítulo 4

6.2.1 Demostración del Teorema 2

Por el Lema 1 para demostrar el Teorema 1 es suficiente ver que para cada conjunto fijo I la función objetivo empírica (4.2) converge a la función objetivo teórica (4.1) c.s.

Observemos que podemos escribir a las funciones objetivo $h_n(I)$ y $h(I)$ del siguiente modo,

$$\begin{aligned}
h_n(I) &= \frac{1}{n}\sum_{j=1}^{n}\left(g\left(\mathbf{X}_j,\beta_n\right) - g\left(\hat{\mathbf{X}}_j(I),\beta_n\right)\right)^2 \\
&= \frac{1}{n}\sum_{j=1}^{n} g^2\left(\mathbf{X}_j,\beta_n\right) - \frac{2}{n}\sum_{j=1}^{n} g\left(\mathbf{X}_j,\beta_n\right) g\left(\hat{\mathbf{X}}_j(I),\beta_n\right) + \frac{1}{n}\sum_{j=1}^{n} g^2\left(\hat{\mathbf{X}}_j(I),\beta_n\right)
\end{aligned}$$

y

$$\begin{aligned}
h(I) &= E\left(\left(g\left(\mathbf{X},\beta_0\right) - g\left(\mathbf{Z}(I),\beta_0\right)\right)^2\right) \\
&= E\left(g^2\left(\mathbf{X},\beta_0\right)\right) - 2E\left(g\left(\mathbf{X},\beta_0\right) g\left(\mathbf{Z}(I),\beta_0\right)\right) + E\left(g^2\left(\mathbf{Z}(I),\beta_0\right)\right).
\end{aligned}$$

Para probarlo veamos que

$$\frac{1}{n}\sum_{j=1}^{n} g^2\left(\mathbf{X}_j,\beta_n\right) \to_{c.s.} E\left(g^2\left(\mathbf{X},\beta_0\right)\right), \tag{6.9}$$

$$\frac{1}{n}\sum_{j=1}^{n} g^2\left(\hat{\mathbf{X}}_j(I),\beta_n\right) \to_{c.s.} E\left(g^2\left(\mathbf{Z}(I),\beta_0\right)\right) \tag{6.10}$$

y

$$\frac{1}{n}\sum_{j=1}^{n} g\left(\mathbf{X}_j,\beta_n\right) g\left(\hat{\mathbf{X}}_j(I),\beta_n\right) \to_{c.s.} E\left(g\left(\mathbf{X},\beta_0\right) g\left(\mathbf{Z}(I),\beta_0\right)\right). \tag{6.11}$$

Primero veamos (6.9). Observemos que,

$$\begin{aligned} & \frac{1}{n}\sum_{j=1}^{n} g^2\left(\mathbf{X}_j,\beta_n\right) \\ = \; & \frac{1}{n}\sum_{j=1}^{n}\left(g^2\left(\mathbf{X}_j,\beta_n\right) - g^2\left(\mathbf{X}_j,\beta_0\right)\right) + \\ & \frac{1}{n}\sum_{j=1}^{n} g^2\left(\mathbf{X}_j,\beta_0\right). \end{aligned} \tag{6.12}$$

Por un lado, la convergencia c.s. a cero de (6.12) es una consecuencia de **HR1**, es decir, como $\|\beta_n - \beta_0\| \to_{c.s.} 0$ y g es una función continua, entonces (6.12) converge c.s a cero. Por otro lado, $\left\{g\left(\mathbf{X}_j,\beta_0\right),\ \text{para } j = 1,\ldots,n\right\}$ son v.a.i.i.d. con segundo momento finito por **HR2**, luego por la Ley Fuerte de los Grandes Números tenemos que

$$\frac{1}{n}\sum_{j=1}^{n} g^2\left(\mathbf{X}_j,\beta_0\right) \to_{c.s.} E\left(g^2\left(\mathbf{X},\beta_0\right)\right).$$

En segundo lugar, veamos (6.10), para eso consideramos los vectores aleatorios no observables $\mathbf{Z}_1(I),\ldots,\mathbf{Z}_n(I)$ definidos en (6.5). Notemos que

$$\begin{aligned} & \frac{1}{n}\sum_{j=1}^{n} g^2\left(\hat{\mathbf{X}}_j(I),\beta_n\right) \\ = \; & \frac{1}{n}\sum_{j=1}^{n}\left(g^2\left(\hat{\mathbf{X}}_j(I),\beta_n\right) - g^2\left(\hat{\mathbf{X}}_j(I),\beta_0\right)\right) + \qquad (6.13) \\ & \frac{1}{n}\sum_{j=1}^{n}\left(g^2\left(\hat{\mathbf{X}}_j(I),\beta_0\right) - g^2\left(\mathbf{Z}_j(I),\beta_0\right)\right) + \qquad (6.14) \\ & \frac{1}{n}\sum_{j=1}^{n} g^2\left(\mathbf{Z}_j(I),\beta_0\right). \end{aligned}$$

Por **HR1** (6.13) converge a cero c.s. Debido a **H1** y a la continuidad de g (ver **HR1**), (6.14) converge a cero c.s. Además $\left\{g\left(\mathbf{Z}_j(I),\beta_0\right) \text{ para } j=1,\ldots,n\right\}$ son v.a.i.i.d. con segundo momento finito pues $E\left(g^2\left(\mathbf{Z}(I),\beta_0\right)\right) < \infty$ por **HR2**, luego por la Ley Fuerte de los Grandes Números tenemos que

$$\frac{1}{n}\sum_{j=1}^{n} g^2\left(\mathbf{Z}_j(I),\beta_0\right) \rightarrow_{c.s.} E\left(g^2\left(\mathbf{Z}(I),\beta_0\right)\right).$$

En tercer lugar, veamos (6.11) utilizando ideas similares.

$$\begin{aligned}
& \frac{1}{n}\sum_{j=1}^{n} g\left(\mathbf{X}_j,\beta_n\right) g\left(\hat{\mathbf{X}}_j(I),\beta_n\right) \\
= \; & \frac{1}{n}\sum_{j=1}^{n}\left(g\left(\mathbf{X}_j,\beta_n\right) g\left(\hat{\mathbf{X}}_j(I),\beta_n\right) - g\left(\mathbf{X}_j,\beta_0\right) g\left(\hat{\mathbf{X}}_j(I),\beta_0\right)\right) + \qquad (6.15) \\
& \frac{1}{n}\sum_{j=1}^{n}\left(g\left(\mathbf{X}_j,\beta_0\right) g\left(\hat{\mathbf{X}}_j(I),\beta_0\right) - g\left(\mathbf{X}_j,\beta_0\right) g\left(\mathbf{Z}_j(I),\beta_0\right)\right) + \qquad (6.16) \\
& \frac{1}{n}\sum_{j=1}^{n} g\left(\mathbf{X}_j,\beta_0\right) g\left(\mathbf{Z}_j(I),\beta_0\right).
\end{aligned}$$

Por **HR1** (6.15) converge c.s. a cero. Debido a **H1** y a la continuidad de g (ver **HR1**) (6.16) converge c.s. a cero. Observemos que

$$E\left(\left|g\left(\mathbf{X},\beta_0\right) g\left(\mathbf{Z}(I),\beta_0\right)\right|\right) \leq E\left(g^2\left(\mathbf{X},\beta_0\right)\right) + E\left(g^2\left(\mathbf{Z}(I),\beta_0\right)\right) < \infty$$

por **HR2**. Luego $\left\{g\left(\mathbf{X}_j,\beta_0\right) g\left(\mathbf{Z}_j(I),\beta_0\right), \text{ para } j=1,\ldots,n\right\}$ son v.a.i.i.d. con primer momento finito, entonces por la Ley Fuerte de los Grandes Números tenemos que

$$\frac{1}{n}\sum_{j=1}^{n} g\left(\mathbf{X}_j,\beta_0\right) g\left(\mathbf{Z}_j(I),\beta_0\right) \rightarrow_{c.s.} E\left(g\left(\mathbf{X},\beta_0\right) g\left(\mathbf{Z}(I),\beta_0\right)\right).$$

Luego queda demostrado el Teorema 2.

6.2.2 Demostración del Teorema 4

Por el Lema 1 para demostrar el Teorema 4, alcanza con ver que para cada conjunto fijo I la función objetivo empírica (4.6) converge c.s. a (4.5). Para simplificar la notación demostramos, sin pérdida de generalidad, el caso de una sola componente (i.e., l=1). Llamamos $\alpha = \alpha_1$.

Luego tenemos,

$$\begin{aligned}
h_n(I) &= \frac{1}{n}\sum_{j=1}^{n}\left(\alpha'^n\mathbf{X}_j - \alpha'^n\hat{\mathbf{X}}_j(I)\right)^2 \\
&= \frac{1}{n}\sum_{j=1}^{n}\left(\sum_{i\notin I}\alpha^n[i]\left(X_j[i]-\hat{X}_j(I)[i]\right)\right)^2 \\
&= \frac{1}{n}\sum_{j=1}^{n}\left(\sum_{i\notin I}(\alpha^n[i])^2\left(X_j[i]-\hat{X}_j(I)[i]\right)^2 + 2\sum_{i,k\notin I,i<k}\alpha^n[i]\alpha^n[k]\left(X_j[i]-\hat{X}_j(I)[i]\right)\left(X_j[k]-\hat{X}_j(I)[k]\right)\right) \\
&= \sum_{i\notin I}(\alpha^n[i])^2\frac{1}{n}\sum_{j=1}^{n}\left(X_j[i]-\hat{X}_j(I)[i]\right)^2 + \\
&\quad 2\sum_{i,k\notin I,i<k}\alpha^n[i]\alpha^n[k]\frac{1}{n}\sum_{j=1}^{n}\left(X_j[i]-\hat{X}_j(I)[i]\right)\left(X_j[k]-\hat{X}_j(I)[k]\right)
\end{aligned}$$

y

$$\begin{aligned}
h(I) &= E\left(\left(\alpha'\mathbf{X}-\alpha'\mathbf{Z}(I)\right)^2\right) \\
&= E\left(\left(\sum_{i\notin I}\alpha[i]\left(X[i]-Z(I)[i]\right)\right)^2\right) \\
&= E\left(\sum_{i\notin I}(\alpha[i])^2\left(X[i]-Z(I)[i]\right)^2 + 2\sum_{i,k\notin I,i<k}\alpha[i]\alpha[k]\left(X[i]-Z(I)[i]\right)\left(X[k]-Z(I)[k]\right)\right) \\
&= \sum_{i\notin I}(\alpha[i])^2 E\left(\left(X[i]-Z(I)[i]\right)^2\right) + \\
&\quad 2\sum_{i,k\notin I,i<k}\alpha[i]\alpha[k]E\left(\left(X[i]-Z(I)[i]\right)\left(X[k]-Z(I)[k]\right)\right).
\end{aligned}$$

Como ya mencionamos al demostrar el Teorema 1, Dauxois et al. (1982) demostraron que bajo **HP1**, si

$$\sup_{\|u\|=1}\left\|\left(\hat{\Sigma}_n-\Sigma\right)(u)\right\| \to_{c.s.} 0,$$

entonces

$$\left\|\alpha_k^n(P_n)-\alpha_k(P)\right\| \to_{c.s.} 0 \text{ para todo } 1\le k\le p,$$

donde α_k^n (respec. α_k) son los autovectores de la matriz de covarianza empírica asociada a P_n ($\hat{\Sigma}_n$) (respec. Σ es la matriz de covarianza asociada a P).

De este modo completaremos la demostración si vemos que

$$\frac{1}{n}\sum_{j=1}^{n}\left(X_j[i]-\hat{X}_j(I)[i]\right)^2 \rightarrow_{c.s.} E\left((X[i]-Z(I)[i])^2\right) \tag{6.17}$$

y

$$\frac{1}{n}\sum_{j=1}^{n}\left(X_j[i]-\hat{X}_j(I)[i]\right)\left(X_j[k]-\hat{X}_j(I)[k]\right) \rightarrow_{c.s.} E\left((X[i]-Z(I)[i])\,(X[k]-Z(I)[k])\right). \tag{6.18}$$

Primero demostramos (6.17).
Recordando la definición (6.5) tenemos,

$$\begin{aligned}
\frac{1}{n}\sum_{j=1}^{n}\left(X_j[i]-\hat{X}_j(I)[i]\right)^2 &= \frac{1}{n}\sum_{j=1}^{n}\left(X_j[i]-Z_j(I)[i]+Z_j(I)[i]-\hat{X}_j(I)[i]\right)^2 \\
&= \frac{1}{n}\sum_{j=1}^{n}\left(X_j[i]-Z_j(I)[i]\right)^2 + \qquad (6.19)\\
&\quad \frac{1}{n}\sum_{j=1}^{n}\left(Z_j(I)[i]-\hat{X}_j(I)[i]\right)^2 + \qquad (6.20)\\
&\quad \frac{2}{n}\sum_{j=1}^{n}\left(X_j[i]-Z_j(I)[i]\right)\left(Z_j(I)[i]-\hat{X}_j(I)[i]\right). \qquad (6.21)
\end{aligned}$$

Por **HP2** tenemos que $\left\{X_j[i]-Z_j(I)[i],\ \text{para } j=1,\ldots,n\right\}$ son v.a.i.i.d. con segundo momento finito, luego por la Ley Fuerte de los Grandes Números, tenemos que (6.19) converge a la esperanza, es decir

$$\frac{1}{n}\sum_{j=1}^{n}\left(X_j[i]-Z_j(I)[i]\right)^2 \rightarrow_{c.s.} E\left((X[i]-Z(I)[i])^2\right).$$

La convergencia c.s. a cero del término (6.20) es una consecuencia directa de **H1**.
Para demostrar que el término (6.21) converge a cero c.s. consideramos la desigualdad de Cauchy–Schwarz. Tenemos

$$\left|\frac{2}{n}\sum_{j=1}^{n}\left(X_j[i]-Z_j(I)[i]\right)\left(Z_j(I)[i]-\hat{X}_j(I)[i]\right)\right| \leq$$

$$2\left(\frac{1}{n}\sum_{j=1}^{n}\left(X_j[i]-Z_j(I)[i]\right)^2\right)^{1/2}\left(\frac{1}{n}\sum_{j=1}^{n}\left(Z_j(I)[i]-\hat{X}_j(I)[i]\right)^2\right)^{1/2}, \tag{6.22}$$

el término de la izquierda en (6.22) por la Ley Fuerte de los Grandes Números y **HP2** converge c.s. a $E\left((X[i] - Z(I)[i])^2\right)$ y el término de la derecha en (6.22) converge c.s. a cero por **H1**.

En segundo lugar demostramos (6.18) utilizando una idea similar.

$$\begin{aligned}
&\frac{1}{n}\sum_{j=1}^{n}\left(X_j[i] - \hat{X}_j(I)[i]\right)\left(X_j[k] - \hat{X}_j(I)[k]\right) = \\
= \;& \frac{1}{n}\sum_{j=1}^{n}\left(X_j[i] - Z_j(I)[i] + Z_j(I)[i] - \hat{X}_j(I)[i]\right)\left(X_j[k] - Z_j(I)[k] + Z_j(I)[k] - \hat{X}_j(I)[k]\right) \\
= \;& \frac{1}{n}\sum_{j=1}^{n}\left(X_j[i] - Z_j(I)[i]\right)\left(X_j[k] - Z_j(I)[k]\right) + \qquad (6.23) \\
& \frac{1}{n}\sum_{j=1}^{n}\left(X_j[i] - Z_j(I)[i]\right)\left(Z_j(I)[k] - \hat{X}_j(I)[k]\right) + \qquad (6.24) \\
& \frac{1}{n}\sum_{j=1}^{n}\left(Z_j(I)[i] - \hat{X}_j(I)[i]\right)\left(X_j[k] - Z_j(I)[k]\right) + \qquad (6.25) \\
& \frac{1}{n}\sum_{j=1}^{n}\left(Z_j(I)[i] - \hat{X}_j(I)[i]\right)\left(Z_j(I)[k] - \hat{X}_j(I)[k]\right). \qquad (6.26)
\end{aligned}$$

Para $i \notin I$, como $\left\{X_j[i] - Z_j(I)[i], \text{ para } j = 1, \ldots, n\right\}$ son v.a.i.i.d. con segundo momento finito (**HP2**), tenemos que, $\left\{\left(X_j[i] - Z_j(I)[i]\right)\left(X_j[k] - Z_j(I)[k]\right), \text{ para } j = 1, \ldots, n\right\}$ son v.a.i.i.d. con primer momento finito. En consecuencia por la Ley Fuerte de los Grandes Números resulta que (6.23) converge a la esperanza, es decir,

$$\frac{1}{n}\sum_{j=1}^{n}\left(X_j[i] - Z_j(I)[i]\right)\left(X_j[k] - Z_j(I)[k]\right) \to_{c.s.} E\left(\left(X[i] - Z(I)[i]\right)\left(X[k] - Z(I)[k]\right)\right).$$

Por la desigualdad de Cauchy–Schwarz, sabemos que,

$$\begin{aligned}
&\left|\frac{1}{n}\sum_{j=1}^{n}\left(X_j[i] - Z_j(I)[i]\right)\left(Z_j(I)[k] - \hat{X}_j(I)[k]\right)\right| \leq \\
&\left(\frac{1}{n}\sum_{j=1}^{n}\left(X_j[i] - Z_j(I)[i]\right)^2\right)^{1/2}\left(\frac{1}{n}\sum_{j=1}^{n}\left(Z_j(I)[k] - \hat{X}_j(I)[k]\right)^2\right)^{1/2}. \qquad (6.27)
\end{aligned}$$

el término de la izquierda en (6.27) por la Ley Fuerte de los Grandes Números y por **HP2** converge c.s. a $E\left((X[i] - Z(I)[i])^2\right)$ y el término de la derecha en (6.27) converge c.s. a cero por **H1**. Por consiguiente (6.24) converge c.s. a cero.

Utilizando los mismos argumentos tenemos que el término (6.25) converge c.s. a cero.

Para demostrar que (6.26) converge c.s. a cero, utilizamos la desigualdad de Cauchy–Schwarz,

$$\left|\frac{1}{n}\sum_{j=1}^{n}\left(Z_j(I)[i]-\hat{X}_j(I)[i]\right)\left(Z_j(I)[k]-\hat{X}_j(I)[k]\right)\right| \leq$$
$$\left(\frac{1}{n}\sum_{j=1}^{n}\left(Z_j(I)[i]-\hat{X}_j(I)[i]\right)^2\right)^{1/2}\left(\frac{1}{n}\sum_{j=1}^{n}\left(Z_j(I)[k]-\hat{X}_j(I)[k]\right)^2\right)^{1/2}.$$

Ambos términos convergen a cero por **H1**.

Por tanto, demostramos el Teorema 4.

6.3 Demostraciones Correspondientes al Capítulo 5

6.3.1 Demostración del Teorema 6

Por el Lema 1 para demostrar el Teorema 6 es suficiente mostrar que (5.5) converge c.s. a (5.4) para todo $I \in \mathcal{I}_d$. Para probarlo alcanza con ver que para cada $k = 1, \ldots, K$,

$$\frac{1}{n}\sum_{j=1}^{n} I_{\{g_n(X_j)=k\}} I_{\{g_n(\hat{X}_j(I))=k\}} \rightarrow_{c.s.} P\left(g(X)=k, g\left(Z(I)\right)=k\right). \tag{6.28}$$

Definimos los procesos no observables $Z_1(I), \ldots, Z_n(I)$, donde

$$Z_j(I)(t) = E\left(X_j(t)|\mathbf{f}\left(I, X_j\right)\right), \tag{6.29}$$

y denotamos $Q_n^*(I)$ su distribución empírica.

La convergencia (6.28) quedará demostrada si probamos que para cada k fijo

$$\frac{1}{n}\sum_{j=1}^{n} I_{\{g_n(X_j)=k\}} I_{\{g_n(Z_j(I))=k\}} \rightarrow_{c.s.} P\left(g(X)=k, g(Z(I))=k\right) \tag{6.30}$$

y

$$\frac{1}{n}\sum_{j=1}^{n} I_{\{g_n(X_j)=k\}}\left(I_{\{g_n(\hat{X}_j(I))=k\}} - I_{\{g_n(Z_j(I))=k\}}\right) \rightarrow_{c.s.} 0. \tag{6.31}$$

Para demostrar (6.30) veamos que

$$\frac{1}{n}\sum_{j=1}^{n} I_{\{g_n(X_j)=k\}} I_{\{g_n(Z_j(I))=k\}} - I_{\{g(X_j)=k\}} I_{\{g(Z_j(I))=k\}} \rightarrow_{c.s.} 0 \tag{6.32}$$

y

$$\frac{1}{n}\sum_{j=1}^{n} I_{\{g(X_j)=k\}} I_{\{g(Z_j(I))=k\}} \to_{c.s.} P\left(g(X)=k, g(Z(I))=k\right). \tag{6.33}$$

La convergencia (6.33) es una consecuencia directa de la Ley Fuerte de los Grandes Números.

Para demostrar (6.32) observamos que el lado izquierdo está mayorado por

$$\frac{1}{n}\sum_{\{X_j\in C(\epsilon,r)\}\cap\{Z_j(I)\in C(\epsilon,r)\}} \left| I_{\{g_n(X_j)=k\}} I_{\{g_n(Z_j(I))=k\}} - I_{\{g(X_j)=k\}} I_{\{g(Z_j(I))=k\}} \right| +$$
$$\frac{1}{n}\sum_{\{X_j\notin C(\epsilon,r)\}\cup\{Z_j(I)\notin C(\epsilon,r)\}} \left| I_{\{g_n(X_j)=k\}} I_{\{g_n(Z_j(I))=k\}} - I_{\{g(X_j)=k\}} I_{\{g(Z_j(I))=k\}} \right|,$$

donde $C(\epsilon, r)$ está dado en la hipótesis **HfC1**(a). El primer término está mayorado por

$$\frac{1}{n}\sum_{\{X_j\in C(\epsilon,r)\}\cap\{Z_j(I)\in C(\epsilon,r)\}} \left(\left| I_{\{g_n(X_j)=k\}} - I_{\{g(X_j)=k\}} \right| + \left| I_{\{g_n(Z_j(I))=k\}} - I_{\{g(Z_j(I))=k\}} \right| \right),$$

y converge a cero c.s. para cualquier ϵ, r positivo por **HfC1**(a), mientras que el segundo término está dominado por

$$\frac{1}{n}\sharp\left\{1\le j\le n : \left\{X_j\notin C(\epsilon,r)\right\}\cup\left\{Z_j(I)\notin C(\epsilon,r)\right\}\right\}$$

que por la Ley Fuerte de los Grandes Números converge c.s. a $P(\{X \notin C(\epsilon, r)\} \cup \{Z(I) \notin C(\epsilon, r)\})$. Eligiendo ϵ y r adecuadamente esta probabilidad es arbitrariamente pequeña, por ende, (6.32) queda demostrado. De este modo (6.30) queda demostrado.

Para demostrar (6.31), veamos que

$$\frac{1}{n}\sharp\left\{1\le j\le n : g_n\left(\hat{X}_j(I)\right)=k, g_n\left(Z_j(I)\right)\neq k, g_n\left(X_j\right)=k\right\} \to_{c.s.} 0, \tag{6.34}$$

y

$$\frac{1}{n}\sharp\left\{1\le j\le n : g_n\left(\hat{X}_j(I)\right)\neq k, g_n\left(Z_j(I)\right)=k, g_n\left(X_j\right)=k\right\} \to_{c.s.} 0. \tag{6.35}$$

Definimos los conjuntos B, C y D del siguiente modo:

$$B=\left\{\omega\in\Omega : \left\|Z_j(I)-\hat{X}_j(I)\right\|_{L^2[a,b]}\to 0 \text{ cuando } n\to\infty \text{ para todo } j\right\},$$

$$C_j=\left\{\omega\in\Omega : d\left(X_j,\partial G_k^n\right)-d\left(X_j,\partial G_k\right)\to 0 \text{ cuando } n\to\infty\right\} \text{ y } C=\cap_{j=1}^{\infty}C_j,$$

$$D = \{\omega \in \Omega : d(\partial G_k^n, \partial G_k) \to 0 \text{ cuando } n \to \infty\}.$$

Por **HfC1**(b) y **Hf1** tenemos que $P(B \cap C) = 1$.
Para ver que la $P(D) = 1$, observemos que,

$$d(\partial G_k^n, \partial G_k) \to_{c.s.} 0.$$

En efecto, si no fuera así, existiría $\epsilon > 0$ y una sucesión $x_n \in \partial G_k^n$ tal que $d(x_n, \partial G_k) > \epsilon$ y $P\left(X_j \in B\left(x_n, \epsilon/4\right)\right) > 0$ por **HfC3**. Es decir,

$$P\left(d\left(X_j, x_n\right) < \epsilon/4, d\left(x_n, \partial G_k\right) > \epsilon\right) > 0.$$

Como

$$\epsilon < d\left(x_n, \partial G_k\right) \leq d\left(x_n, X_j\right) + d\left(X_j, \partial G_k\right),$$

$$\text{si } d\left(X_j, x_n\right) < \frac{\epsilon}{4} \text{ entonces } d\left(X_j, \partial G_k\right) > \frac{3\epsilon}{4}.$$

A su vez como $x_n \in \partial G_k^n$ concluimos que

$$P\left(d\left(X_j, \partial G_k^n\right) < \epsilon/4, d\left(X_j, \partial G_k\right) > 3\epsilon/4\right) > 0.$$

Lo que contradice **HfC1**(b).

Por lo tanto, $P(B \cap C \cap D) = 1$. En consecuencia, dado $\delta > 0$ y $\omega \in B \cap C \cap D$, existe $n_0(\omega, \delta)$ tal que para todo $n \geq n_0(\omega, \delta)$,

$$max_{j=1,\ldots,n} \left\|Z_j(I) - \hat{X}_j(I)\right\|_{L^2[a,b]} \leq \delta/2.$$

Dado $\omega \in B \cap C \cap D$, $\delta > 0$ y $n \geq n_0(\omega, \delta)$ veamos que:

$$\begin{aligned}\left\{1 \leq j \leq n : g_n\left(\hat{X}_j(I)\right) = k, g_n\left(Z_j(I)\right) \neq k, g_n\left(X_j\right) = k\right\} &\subseteq \left\{1 \leq j \leq n : d\left(\hat{X}_j(I), \partial G_k^n\right) < \delta\right\} \\ &\subseteq \left\{1 \leq j \leq n : d\left(Z_j(I), \partial G_k\right) < 2\delta\right\}.\end{aligned}$$

Para demostrar la primera inclusión observemos que para cada j fijo tenemos,

$$\text{si } d\left(\hat{X}_j(I), \partial G_k^n\right) > \delta, \text{ entonces } \overline{B\left(\hat{X}_j(I), \delta/2\right)} \cap \partial G_k^n = \emptyset,$$

además como

$$\left\|Z_j(I) - \hat{X}_j(I)\right\|_{L^2[a,b]} \leq \delta/2, \text{ entonces } Z_j(I) \in B\left(\hat{X}_j(I), \delta/2\right),$$

luego, se contradice $g_n\left(\hat{X}_j(I)\right) = k, g_n\left(Z_j(I)\right) \neq k$.

Para demostrar la segunda inclusión consideremos $n_1(\omega) \geq n_0(\omega, \delta)$ tal que $d\left(\partial G_k^{n_1}, \partial G_k\right) < \delta/2$, entonces obtenemos que para cada j

$$d\left(Z_j(I), \partial G_k\right) \leq d\left(Z_j(I), \hat{X}_j(I)\right) + d\left(\hat{X}_j(I), \partial G_k^{n_1}\right) + d\left(\partial G_k^{n_1}, \partial G_k\right) < \frac{\delta}{2} + \delta + \frac{\delta}{2} = 2\delta.$$

Luego, demostramos que con probabilidad uno

$$\begin{aligned} \frac{1}{n}\sharp\left\{1 \leq j \leq n : g_n\left(\hat{X}_j(I)\right) = k, g_n\left(Z_j(I)\right) \neq k, g_n\left(X_j\right) = k\right\} &\leq \frac{1}{n}\sharp\left\{1 \leq j \leq n : d\left(Z_j(I), \partial G_k\right) < 2\delta\right\} \\ &= \frac{1}{n}\sum_{j=1}^{n} I_{\left\{d\left(Z_j(I), \partial G_k\right) < 2\delta\right\}}. \end{aligned}$$

Por la Ley Fuerte de los Grandes Números tenemos que

$$\frac{1}{n}\sum_{j=1}^{n} I_{\left\{d\left(Z_j(I), \partial G_k\right) < 2\delta\right\}} \rightarrow_{c.s.} P\left(d\left(Z_j(I), \partial G_k\right) < 2\delta\right)$$

y por **HfC2** tenemos que $P\left(d\left(Z_j(I), \partial G_k\right) < 2\delta\right) \rightarrow 0$ cuando $\delta \rightarrow 0$. Conjuntamente demostramos (6.34).

La demostración de (6.35) es completamente análoga.

Por consiguiente queda demostrado (6.31), concluyendo la demostración del Teorema 6.

6.3.2 Demostración del Teorema 7

Por el Lema 1, para demostrar el Teorema 7, es suficiente ver que (5.11) converge c.s. a (5.9). Para simplificar la notación consideramos, sin pérdida de generalidad, una sola componente principal, i.e. $l = 1$. Denotamos α a la función de pesos asociada a esta componente. De este modo tenemos,

$$\begin{aligned} h_n(I) &= \frac{1}{n}\sum_{j=1}^{n}\left(U_1^j - U_1^j(I)\right)^2 \\ &= \frac{1}{n}\sum_{j=1}^{n}\left(\left\langle \alpha^n, X_j\right\rangle_{L^2[a,b]} - \left\langle \alpha^n, \hat{X}_j(I)\right\rangle_{L^2[a,b]}\right)^2 \\ &= \frac{1}{n}\sum_{j=1}^{n}\left(\int_a^b \alpha^n(t)\left(X_j(t) - \hat{X}_j(I)(t)\right)dt\right)^2, \end{aligned}$$

y

$$\begin{aligned} h(I) &= E\left((U_1 - U_1(I))^2\right) \\ &= E\left(\left(\langle \alpha, X\rangle_{L^2[a,b]} - \langle \alpha, Z(I)\rangle_{L^2[a,b]}\right)^2\right) \\ &= E\left(\left(\int_a^b \alpha(t)(X(t) - Z(I)(t))dt\right)^2\right). \end{aligned}$$

Para la demostración consideramos los procesos no observables definidos anteriormente en (6.29).

Observemos que la función objetivo empírica puede ser escrita como

$$\begin{aligned} h_n(I) &= \frac{1}{n}\sum_{j=1}^{n}\left(\int_a^b \alpha^n(t)\left(X_j(t) - Z_j(I)(t) + Z_j(I)(t) - \hat{X}_j(I)(t)\right)dt\right)^2 \\ &= \frac{1}{n}\sum_{j=1}^{n}\left(\int_a^b \alpha^n(t)\left(X_j(t) - Z_j(I)(t)\right)dt\right)^2 + \qquad (6.36) \\ &\quad \frac{1}{n}\sum_{j=1}^{n}\left(\int_a^b \alpha^n(t)\left(Z_j(I)(t) - \hat{X}_j(I)(t)\right)dt\right)^2 + \qquad (6.37) \\ &\quad \frac{2}{n}\sum_{j=1}^{n}\left(\int_a^b \alpha^n(t)\left(X_j(t) - Z_j(I)(t)\right)dt\right)\left(\int_a^b \alpha^n(t)\left(Z_j(I)(t) - \hat{X}_j(I)(t)\right)dt\right). \qquad (6.38) \end{aligned}$$

En primer lugar demostramos que el término (6.36) converge a $h(I)$, para eso observamos que,

$$\begin{aligned} &\frac{1}{n}\sum_{j=1}^{n}\left(\int_a^b \alpha^n(t)\left(X_j(t) - Z_j(I)(t)\right)dt\right)^2 \\ = &\frac{1}{n}\sum_{j=1}^{n}\left(\int_a^b (\alpha^n(t) - \alpha(t) + \alpha(t))\left(X_j(t) - Z_j(I)(t)\right)dt\right)^2 \\ = &\frac{1}{n}\sum_{j=1}^{n}\left(\int_a^b (\alpha^n(t) - \alpha(t))\left(X_j(t) - Z_j(I)(t)\right)dt\right)^2 + \qquad (6.39) \\ &\frac{1}{n}\sum_{j=1}^{n}\left(\int_a^b \alpha(t)\left(X_j(t) - Z_j(I)(t)\right)dt\right)^2 + \qquad (6.40) \\ &\frac{2}{n}\sum_{j=1}^{n}\left(\int_a^b (\alpha^n(t) - \alpha(t))\left(X_j(t) - Z_j(I)(t)\right)dt\right)\left(\int_a^b \alpha(t)\left(X_j(t) - Z_j(I)(t)\right)dt\right). \qquad (6.41) \end{aligned}$$

Veamos que (6.39) converge c.s. a cero. Por la desigualdad de Cauchy–Schwarz tenemos que

$$\begin{aligned} \frac{1}{n}\sum_{j=1}^{n}\left(\int_a^b (\alpha^n(t) - \alpha(t))\left(X_j(t) - Z_j(I)(t)\right)dt\right)^2 &\leq \frac{1}{n}\sum_{j=1}^{n}\|\alpha^n - \alpha\|^2_{L^2[a,b]}\left\|X_j - Z_j(I)\right\|^2_{L^2[a,b]} \\ &= \|\alpha^n - \alpha\|^2_{L^2[a,b]}\,\frac{1}{n}\sum_{j=1}^{n}\left\|X_j - Z_j(I)\right\|^2_{L^2[a,b]}. \end{aligned}$$

Por un lado, como $\left\{\left\|X_j - Z_j(I)\right\|_{L^2[a,b]},\ \text{para } j = 1,\ldots,n\right\}$ son v.a.i.i.d. con segundo momento finito por **Hf2** y por la Ley Fuerte de los Grandes Números tenemos que,

$$\frac{1}{n}\sum_{j=1}^{n}\left\|X_j - Z_j(I)\right\|^2_{L^2[a,b]} \to_{c.s.} E\left(\|X - Z(I)\|\right).$$

Por otro lado, por la Ley Fuerte de los Grandes Números y por **HfP1** tenemos que

$$\|\mathbf{\Gamma}_n - \mathbf{\Gamma}\|_{\mathcal{F}} \to_{c.s.} 0,$$

luego de Dauxois et al. (1982) se deriva que,

$$\left\|\alpha_k^n - \alpha_k\right\|_{L^2[a,b]} \to_{c.s.} 0, \quad \text{para todo } k \geq 1,$$

donde α_k^n (respec. α_k) son las autofunciones de $\mathbf{\Gamma}_n$ (respec. $\mathbf{\Gamma}$). En consecuencia demostramos que (6.39) converge c.s. a cero.

Para demostrar que (6.40) converge c.s. a $h(I)$, observemos que por la desigualdad de Cauchy–Schwarz tenemos que

$$\left(\int_a^b \alpha(t)\left(X_j(t) - Z_j(I)(t)\right)dt\right)^2 \leq \|\alpha\|^2_{L^2[a,b]}\left\|X_j - Z_j(I)\right\|^2_{L^2[a,b]},$$

Luego, por **Hf2** tenemos que $\left\{\int_a^b \alpha(t)\left(X_j(t) - Z_j(I)(t)\right)dt,\ \text{para } j = 1,\ldots,n\right\}$ son v.a.i.i.d. con segundo momento finito y por la Ley Fuerte de los Grandes Números tenemos que

$$\frac{1}{n}\sum_{j=1}^{n}\left(\int_a^b \alpha(t)\left(X_j(t) - Z_j(I)(t)\right)dt\right)^2 \to_{c.s.} E\left(\left(\int_a^b \alpha(t)\left(X(t) - Z(I)(t)\right)dt\right)^2\right).$$

Veamos que el término (6.41) converge c.s. a cero. Por la desigualdad de Cauchy–Schwarz tenemos que

$$\begin{aligned}
&\left|\frac{2}{n}\sum_{j=1}^{n}\left(\int_a^b (\alpha^n(t) - \alpha(t))\left(X_j(t) - Z_j(I)(t)\right)dt\right)\left(\int_a^b \alpha(t)\left(X_j(t) - Z_j(I)(t)\right)dt\right)\right| \\
\leq\ & 2\left(\frac{1}{n}\sum_{j=1}^{n}\left(\int_a^b (\alpha^n(t) - \alpha(t))\left(X_j(t) - Z_j(I)(t)\right)dt\right)^2\right)^{1/2}\left(\frac{1}{n}\sum_{j=1}^{n}\left(\int_a^b \alpha(t)\left(X_j(t) - Z_j(I)(t)\right)dt\right)^2\right)^{1/2}.
\end{aligned}$$

Y acabamos de probar que el término de la izquierda converge a cero y que el término de la derecha es acotado, por lo tanto el producto converge a cero.

En segundo lugar veamos que (6.37) converge c.s. a cero. Por la desigualdad de Cauchy–Schwarz tenemos que,

$$\begin{aligned}\frac{1}{n}\sum_{j=1}^{n}\left(\int_a^b \alpha^n(t)\left(Z_j(I)(t)-\hat{X}_j(I)(t)\right)dt\right)^2 &\leq \frac{1}{n}\sum_{j=1}^{n}\|\alpha^n\|^2_{L^2[a,b]}\left\|Z_j(I)-\hat{X}_j(I)\right\|^2_{L^2[a,b]}\\ &= \|\alpha^n\|^2_{L^2[a,b]}\frac{1}{n}\sum_{j=1}^{n}\left\|Z_j(I)-\hat{X}_j(I)\right\|^2_{L^2[a,b]}.\end{aligned}$$

El término $\|\alpha^n\|^2_{L^2[a,b]}$ converge c.s. a $\|\alpha\|^2_{L^2[a,b]}$ por Dauxois et al. (1982) y $\frac{1}{n}\sum_{j=1}^{n}\left\|Z_j(I)-\hat{X}_j(I)\right\|^2_{L^2[a,b]}$ converge c.s. a cero por **Hf1**.

Finalmente veamos que el término (6.38) converge c.s. a cero. Por la desigualdad de Cauchy–Schwarz tenemos que,

$$\begin{aligned}&\left|\frac{2}{n}\sum_{j=1}^{n}\left(\int_a^b \alpha^n(t)\left(X_j(t)-Z_j(I)(t)\right)dt\right)\left(\int_a^b \alpha^n(t)\left(Z_j(I)(t)-\hat{X}_j(I)(t)\right)dt\right)\right|\\ \leq\ & 2\left(\frac{1}{n}\sum_{j=1}^{n}\left(\int_a^b \alpha^n(t)\left(X_j(t)-Z_j(I)(t)\right)dt\right)^2\right)^{1/2}\left(\frac{1}{n}\sum_{j=1}^{n}\left(\int_a^b \alpha^n(t)\left(Z_j(I)(t)-\hat{X}_j(I)(t)\right)dt\right)^2\right)^{1/2}.\end{aligned}$$

Por (6.36) y (6.37) tenemos que el primer término es acotado y el segundo converge c.s. a cero, por lo tanto el producto converge c.s. a cero.

Por tanto, concluimos la demostración del Teorema 7.

6.3.3 Demostración del Teorema 8

Por el Lema 1 para demostrar el Teorema 8 es suficiente ver que (5.18) converge c.s. a (5.17), es decir que,

$$\frac{1}{n}\sum_{j=1}^{n}\left(\int_a^b \beta_n(t)X_j(t)dt-\int_a^b \beta_n(t)\hat{X}_j(I)(t)dt\right)^2 \to_{c.s.} E\left(\left(\int_a^b \beta_0(t)X(t)dt-\int_a^b \beta_0(t)Z(I)(t)dt\right)^2\right).$$

Consideramos los procesos no observables definidos anteriormente en (6.29) y observe-

mos que la función objetivo empírica puede ser escrita como

$$\begin{aligned} h_n(I) &= \frac{1}{n}\sum_{j=1}^{n}\left(\int_a^b \beta_n(t)X_j(t)dt - \int_a^b \beta_n(t)\hat{X}_j(I)(t)dt\right)^2 \\ &= \frac{1}{n}\sum_{j=1}^{n}\left(\int_a^b \beta_n(t)\left(X_j(t) - Z_j(I)(t)\right)dt + \int_a^b \beta_n(t)\left(Z_j(I)(t) - \hat{X}_j(I)(t)\right)dt\right)^2 \\ &= \frac{1}{n}\sum_{j=1}^{n}\left(\int_a^b \beta_n(t)\left(X_j(t) - Z_j(I)(t)\right)dt\right)^2 + \qquad (6.42) \\ &\quad \frac{1}{n}\sum_{j=1}^{n}\left(\int_a^b \beta_n(t)\left(Z_j(I)(t) - \hat{X}_j(I)(t)\right)dt\right)^2 + \qquad (6.43) \\ &\quad \frac{2}{n}\sum_{j=1}^{n}\left(\int_a^b \beta_n(t)\left(X_j(t) - Z_j(I)(t)\right)dt\right)\left(\int_a^b \beta_n(t)\left(Z_j(I)(t) - \hat{X}_j(I)(t)\right)dt\right). \qquad (6.44) \end{aligned}$$

Primero veamos que (6.42) converge a $h(I)$. Para eso escribamos a la ecuación (6.42) del siguiente modo,

$$\begin{aligned} &\frac{1}{n}\sum_{j=1}^{n}\left(\int_a^b \beta_n(t)\left(X_j(t) - Z_j(I)(t)\right)dt\right)^2 \\ = &\frac{1}{n}\sum_{j=1}^{n}\left(\int_a^b (\beta_n(t) - \beta_0(t))\left(X_j(t) - Z_j(I)(t)\right)dt\right)^2 + \qquad (6.45) \\ &\frac{1}{n}\sum_{j=1}^{n}\left(\int_a^b \beta_0(t)\left(X_j(t) - Z_j(I)(t)\right)dt\right)^2 + \qquad (6.46) \\ &\frac{2}{n}\sum_{j=1}^{n}\left(\int_a^b (\beta_n(t) - \beta_0(t))\left(X_j(t) - Z_j(I)(t)\right)dt\right)\left(\int_a^b \beta_0(t)\left(X_j(t) - Z_j(I)(t)\right)dt\right). \qquad (6.47) \end{aligned}$$

Veamos que (6.45) converge a cero. Por la desigualdad de Cauchy–Schwarz tenemos que

$$\begin{aligned} \frac{1}{n}\sum_{j=1}^{n}\left(\int_a^b (\beta_n(t) - \beta_0(t))\left(X_j(t) - Z_j(I)(t)\right)dt\right)^2 &\leq \frac{1}{n}\sum_{j=1}^{n}\|\beta_n - \beta_0\|^2_{L^2[a,b]}\left\|X_j - Z_j(I)\right\|^2_{L^2[a,b]} \\ &= \|\beta_n - \beta_0\|^2_{L^2[a,b]}\frac{1}{n}\sum_{j=1}^{n}\left\|X_j - Z_j(I)\right\|^2_{L^2[a,b]}. \end{aligned}$$

Por un lado, por **HfR1** tenemos que

$$\|\beta_n - \beta_0\|^2_{L^2[a,b]} \to_{c.s.} 0.$$

Por el otro, como $\left\{\left\|X_j - Z_j(I)\right\|^2_{L^2[a,b]} \text{ para } j = 1, \ldots, n\right\}$ son v.a.i.i.d. con segundo momento finito por **Hf2**, por la Ley Fuerte de los Grandes Números,

$$\frac{1}{n}\sum_{j=1}^{n}\left\|X_j - Z_j(I)\right\|^2_{L^2[a,b]} \to_{c.s.} E\left(\|X - Z(I)\|^2_{L^2[a,b]}\right).$$

Por lo tanto el producto converge c.s. a cero.

Veamos que (6.46) converge a $h(I)$. Por la desigualdad de Cauchy–Schwarz tenemos que

$$\left(\int_a^b \beta_0(t)\left(X(t) - Z(I)(t)\right)dt\right)^2 \leq \|\beta_0\|^2_{L^2[a,b]}\,\|X - Z(I)\|^2_{L^2[a,b]},$$

entonces,

$$E\left(\left(\int_a^b \beta_0(t)\left(X(t) - Z(I)(t)\right)dt\right)^2\right) \leq \|\beta_0\|^2_{L^2[a,b]}\,E\left(\|X - Z(I)\|^2_{L^2[a,b]}\right).$$

Por **Hf2** $E\left(\|X - Z(I)\|^2_{L^2[a,b]}\right) < \infty$, por tanto, $\left\{\int_a^b \beta_0(t)\left(X_j(t) - Z_j(I)(t)\right)dt \text{ para } j = 1, \ldots, n\right\}$ son v.a.i.i.d. con segundo momento finito. En consecuencia por la Ley Fuerte de los Grandes Números

$$\frac{1}{n}\sum_{j=1}^{n}\left(\int_a^b \beta_0(t)\left(X_j(t) - Z_j(I)(t)\right)dt\right)^2 \to_{c.s.} E\left(\left(\int_a^b \beta_0(t)X(t)dt - \int_a^b \beta_0(t)Z(I)(t)dt\right)^2\right) = h(I).$$

Para demostrar que (6.47) converge c.s. a cero, observemos que por la desigualdad de Cauchy–Schwarz tenemos que

$$\begin{aligned}
&\left|\frac{2}{n}\sum_{j=1}^{n}\left(\int_a^b (\beta_n(t) - \beta_0(t))\left(X_j(t) - Z_j(I)(t)\right)dt\right)\left(\int_a^b \beta_0(t)\left(X_j(t) - Z_j(I)(t)\right)dt\right)\right| \\
\leq\ & 2\left(\frac{1}{n}\sum_{j=1}^{n}\left(\int_a^b (\beta_n(t) - \beta_0(t))\left(X_j(t) - Z_j(I)(t)\right)dt\right)^2\right)^{1/2}\left(\frac{1}{n}\sum_{j=1}^{n}\left(\int_a^b \beta_0(t)\left(X_j(t) - Z_j(I)(t)\right)dt\right)^2\right)^{1/2}.
\end{aligned}$$

Acabamos de demostrar que el término de la izquierda converge c.s. a cero y que el término de la derecha es acotado c.s., por lo tanto el producto converge c.s. a cero.

En segundo lugar veamos que (6.43) converge a cero c.s. Por la desigualdad de Cauchy–Schwarz tenemos que

$$\begin{aligned}
\frac{1}{n}\sum_{j=1}^{n}\left(\int_a^b \beta_n(t)\left(Z_j(I)(t) - \hat{X}_j(I)(t)\right)dt\right)^2 &\leq \frac{1}{n}\sum_{j=1}^{n}\|\beta_n\|^2_{L^2[a,b]}\left\|Z_j(I) - \hat{X}_j(I)\right\|^2_{L^2[a,b]} \\
&= \|\beta_n\|^2_{L^2[a,b]}\,\frac{1}{n}\sum_{j=1}^{n}\left\|Z_j(I) - \hat{X}_j(I)\right\|^2_{L^2[a,b]}.
\end{aligned}$$

El término $\|\beta_n\|^2_{L^2[a,b]}$ es finito c.s. pues por **HfR1** converge c.s. a $\|\beta_0\|^2_{L^2[a,b]}$, y el término de la izquierda converge c.s. a cero por **Hf1**. Por consiguiente el producto converge c.s. a cero.

Finalmente veamos que (6.44) también converge c.s. a cero. Aplicando la desigualdad de Cauchy–Schwarz tenemos que

$$\begin{aligned} & \left| \frac{2}{n} \sum_{j=1}^{n} \left(\int_a^b \beta_n(t) \left(X_j(t) - Z_j(I)(t) \right) dt \right) \left(\int_a^b \beta_n(t) \left(Z_j(I)(t) - \hat{X}_j(I)(t) \right) dt \right) \right| \\ \leq \; & 2 \left(\frac{1}{n} \sum_{j=1}^{n} \left(\int_a^b \beta_n(t) \left(X_j(t) - Z_j(I)(t) \right) dt \right)^2 \right)^{1/2} \left(\frac{1}{n} \sum_{j=1}^{n} \left(\int_a^b \beta_n(t) \left(Z_j(I)(t) - \hat{X}_j(I)(t) \right) dt \right)^2 \right)^{1/2}. \end{aligned}$$

El término de la izquierda es la raíz de la ecuación (6.42), que vimos que es finito c.s. y el término de la derecha es la raíz de la ecuación (6.43) que demostramos que converge c.s. a cero, por lo tanto, el producto converge c.s. a cero.

Luego el Teorema 8 quedó probado.

6.3.4 Demostración del Teorema 9

Observemos que el Lema 1 también vale si la convergencia es en probabilidad. Por lo tanto, para demostrar el Teorema 9 veamos que (5.22) converge en probabilidad a (5.21). Recordemos las definiciones de la función objetivo teórica y empírica, y consideramos los procesos no observables definidos anteriormente en (6.29) para escribir a la función objetivo empírica convenientemente.

$$\begin{aligned} h(I) &= E\left(\left\| \int_a^b \beta_0(t,s) X(t) dt - \int_a^b \beta_0(t,s) Z(I)(t) dt \right\|^2_{L^2[c,d]} \right) \\ &= E\left(\int_c^d \left(\int_a^b \beta_0(t,s) \left(X(t) - Z(I)(t) \right) dt \right)^2 ds \right) \end{aligned}$$

y

$$
\begin{aligned}
h_n(I) &= \frac{1}{n}\sum_{j=1}^{n}\int_c^d\left(\int_a^b \beta_n(t,s)X_j(t)dt - \int_a^b \beta_n(t,s)\hat{X}_j(I)(t)dt\right)^2 ds \\
&= \frac{1}{n}\sum_{j=1}^{n}\int_c^d\left(\int_a^b \beta_n(t,s)\left(X_j(t) - Z_j(I)(t) + Z_j(I)(t) - \hat{X}_j(I)(t)\right)dt\right)^2 ds \\
&= \frac{1}{n}\sum_{j=1}^{n}\int_c^d\left(\int_a^b \beta_n(t,s)\left(X_j(t) - Z_j(I)(t)\right) + \beta_n(t,s)\left(Z_j(I)(t) - \hat{X}_j(I)(t)\right)dt\right)^2 ds \\
&= \frac{1}{n}\sum_{j=1}^{n}\int_c^d\left(\int_a^b \beta_n(t,s)\left(X_j(t) - Z_j(I)(t)\right)dt\right)^2 ds + \qquad (6.48) \\
&\quad \frac{1}{n}\sum_{j=1}^{n}\int_c^d\left(\int_a^b \beta_n(t,s)\left(Z_j(I)(t) - \hat{X}_j(I)(t)\right)dt\right)^2 ds + \qquad (6.49) \\
&\quad \frac{2}{n}\sum_{j=1}^{n}\int_c^d\left(\int_a^b \beta_n(t,s)\left(X_j(t) - Z_j(I)(t)\right)dt\right)\left(\int_a^b \beta_n(t,s)\left(Z_j(t) - \hat{X}_j(I)(t)\right)dt\right)ds. \qquad (6.50)
\end{aligned}
$$

Primero veamos que el término (6.48) converge en probabilidad a $h(I)$. Podemos escribir,

$$
\begin{aligned}
&\frac{1}{n}\sum_{j=1}^{n}\int_c^d\left(\int_a^b \beta_n(t,s)\left(X_j(t) - Z_j(I)(t)\right)dt\right)^2 ds \\
= &\frac{1}{n}\sum_{j=1}^{n}\int_c^d\left(\int_a^b (\beta_n(t,s) - \beta_0(t,s))\left(X_j(t) - Z_j(I)(t)\right)dt\right)^2 ds + \qquad (6.51) \\
&\frac{1}{n}\sum_{j=1}^{n}\int_c^d\left(\int_a^b \beta_0(t,s)\left(X_j(t) - Z_j(I)(t)\right)dt\right)^2 ds + \qquad (6.52) \\
&\frac{2}{n}\sum_{j=1}^{n}\int_c^d\left(\int_a^b (\beta_n(t,s) - \beta_0(t,s))\left(X_j(t) - Z_j(I)(t)\right)dt\right)\left(\int_a^b \beta_0(t,s)\left(X_j(t) - Z_j(I)(t)\right)dt\right)ds. \qquad (6.53)
\end{aligned}
$$

Veamos que (6.51) converge en probabilidad a cero. Por la desigualdad de Cauchy–

Schwarz tenemos que

$$\begin{aligned}
& \frac{1}{n}\sum_{j=1}^{n}\int_c^d\left(\int_a^b(\beta_n(t,s)-\beta_0(t,s))\left(X_j(t)dt-Z_j(I)(t)\right)dt\right)^2 ds \\
\leq\ & \frac{1}{n}\sum_{j=1}^{n}\int_c^d \|\beta_n(\cdot,s)-\beta_0(\cdot,s)\|^2_{L^2[a,b]}\left\|X_j-Z_j(I)\right\|^2_{L^2[a,b]} ds \\
=\ & \frac{1}{n}\sum_{j=1}^{n}\|\beta_n-\beta_0\|^2_{L^2([a,b]\times[c,d])}\left\|X_j-Z_j(I)\right\|^2_{L^2[a,b]} \\
=\ & \|\beta_n-\beta_0\|^2_{L^2([a,b]\times[c,d])}\frac{1}{n}\sum_{j=1}^{n}\left\|X_j-Z_j(I)\right\|^2_{L^2[a,b]}.
\end{aligned}$$

Por un lado, $\left\{\left\|X_j-Z_j(I)\right\|^2_{L^2[a,b]} \text{ para } j=1,\ldots,n\right\}$ son v.a.i.i.d. con segundo momento finito por **Hf2** y por la Ley Fuerte de los Grandes Números

$$\frac{1}{n}\sum_{j=1}^{n}\left\|X_j-Z_j(I)\right\|^2_{L^2[a,b]}\to_{c.s.} E\left(\|X-Z(I)\|^2_{L^2[a,b]}\right).$$

Por el otro, por **HfRf1** tenemos que

$$\|\beta_n-\beta_0\|^2_{L^2([a,b]\times[c,d])}\to_{p.} 0.$$

Por lo tanto, el producto converge en probabilidad a cero.

Demostremos que (6.52) converge a $h(I)$ c.s. y por lo tanto converge en probabilidad. Por la desigualdad de Cauchy–Schwarz tenemos que

$$\begin{aligned}
\left|\int_c^d\left(\int_a^b\beta_0(t,s)\left(X(t)-Z(I)(t)\right)dt\right)^2 ds\right| &\leq \|X-Z(I)\|^2_{L^2[a,b]}\int_c^d\|\beta_0(\cdot,s)\|^2_{L^2[a,b]}\,ds \\
&= \|X-Z(I)\|^2_{L^2[a,b]}\|\beta_0\|^2_{L^2([a,b]\times[c,d])}.
\end{aligned}$$

Entonces,

$$E\left(\left|\int_c^d\left(\int_a^b\beta_0(t,s)\left(X(t)-Z(I)(t)\right)dt\right)^2 ds\right|\right) \leq \|\beta_0\|^2_{L^2([a,b]\times[c,d])}E\left(\|X-Z(I)\|^2_{L^2[a,b]}\right),$$

que es finito por **Hf2**. En consecuencia $\left\{\int_c^d\left(\int_a^b\beta_0(t,s)\left(X_j(t)-Z_j(I)(t)\right)dt\right)^2 ds \text{ para } j=1,\ldots,n\right\}$ son v.a.i.i.d. con primer momento finito, luego por la Ley Fuerte de los Grandes Números tenemos que

$$\frac{1}{n}\sum_{j=1}^{n}\int_c^d\left(\int_a^b\beta_0(t,s)\left(X_j(t)-Z_j(I)(t)\right)dt\right)^2 ds\to_{c.s.} E\left(\int_c^d\left(\int_a^b\beta_0(t,s)\left(X(t)-Z(I)(t)\right)dt\right)^2 ds\right)=h(I).$$

Para ver que (6.53) converge en probabilidad a cero, observemos que por la desigualdad de Cauchy–Schwarz tenemos que

$$
\begin{aligned}
& \left| \frac{2}{n} \sum_{j=1}^{n} \int_c^d \left(\int_a^b (\beta_n(t,s) - \beta_0(t,s)) \left(X_j(t) - Z_j(I)(t) \right) dt \right) \left(\int_a^b \beta_0(t,s) \left(X_j(t) - Z_j(I)(t) \right) dt \right) ds \right| \\
\leq & \frac{2}{n} \sum_{j=1}^{n} \int_c^d \|\beta_n(\cdot,s) - \beta_0(\cdot,s)\|_{L^2[a,b]} \left\| X_j - Z_j(I) \right\|_{L^2[a,b]} \|\beta_0(\cdot,s)\|_{L^2[a,b]} \left\| X_j - Z_j(I) \right\|_{L^2[a,b]} ds \\
= & \frac{2}{n} \sum_{j=1}^{n} \left\| X_j - Z_j(I) \right\|^2_{L^2[a,b]} \int_c^d \|\beta_n(\cdot,s) - \beta_0(\cdot,s)\|_{L^2[a,b]} \|\beta_0(\cdot,s)\|_{L^2[a,b]} ds \\
\leq & 2 \frac{1}{n} \sum_{j=1}^{n} \left\| X_j - Z_j(I) \right\|^2_{L^2[a,b]} \left(\int_c^d \|\beta_n(\cdot,s) - \beta_0(\cdot,s)\|^2_{L^2[a,b]} ds \right)^{1/2} \left(\int_c^d \|\beta_0(\cdot,s)\|^2_{L^2[a,b]} ds \right)^{1/2} \\
= & 2 \|\beta_n - \beta_0\|_{L^2([a,b]\times[c,d])} \|\beta_0\|_{L^2([a,b]\times[c,d])} \frac{1}{n} \sum_{j=1}^{n} \left\| X_j - Z_j(I) \right\|^2_{L^2[a,b]}.
\end{aligned}
$$

Como mencionamos anteriormente por **Hf2** y la Ley Fuerte de los Grandes Números tenemos que $\frac{1}{n} \sum_{j=1}^{n} \left\| X_j - Z_j(I) \right\|^2_{L^2[a,b]}$ es finito c.s., $\|\beta_n - \beta_0\|_{L^2([a,b]\times[c,d])}$ converge en probabilidad a cero por **HfRf1** y $\|\beta_0\|_{L^2([a,b]\times[c,d])}$ es finito, por consiguiente el producto converge a cero en probabilidad.

En segundo lugar veamos que (6.49) converge en probabilidad a cero. Por la desigualdad de Cauchy–Schwarz tenemos que

$$
\begin{aligned}
\frac{1}{n} \sum_{j=1}^{n} \int_c^d \left(\int_a^b \beta_n(t,s) \left(Z_j(I)(t) - \hat{X}_j(I)(t) \right) dt \right)^2 ds & \leq \frac{1}{n} \sum_{j=1}^{n} \int_c^d \|\beta_n(\cdot,s)\|^2_{L^2[a,b]} \left\| Z_j(I) - \hat{X}_j(I) \right\|^2_{L^2[a,b]} ds \\
& = \|\beta_n\|^2_{L^2([a,b]\times[c,d])} \frac{1}{n} \sum_{j=1}^{n} \left\| Z_j(I) - \hat{X}_j(I) \right\|^2_{L^2[a,b]}.
\end{aligned}
$$

El término $\|\beta_n\|^2_{L^2([a,b]\times[c,d])}$ converge por **HfRf1** en probabilidad a $\|\beta_0\|^2_{L^2([a,b]\times[c,d])}$, por lo tanto está acotado en probabilidad, y por **Hf1** converge a cero c.s. el término de la derecha. Luego el producto converge a cero en probabilidad.

Finalmente veamos que (6.50) converge en probabilidad a cero. Por la desigualdad de

Cauchy–Schwarz tenemos que

$$
\begin{aligned}
& \left| \frac{2}{n} \sum_{j=1}^{n} \int_c^d \left(\int_a^b \beta_n(t,s) \left(X_j(t) - Z_j(I)(t) \right) dt \right) \left(\int_a^b \beta_n(t,s) \left(Z_j(t) - \hat{X}_j(I)(t) \right) dt \right) ds \right| \\
\leq & \frac{2}{n} \sum_{j=1}^{n} \int_c^d \|\beta_n(\cdot,s)\|_{L^2[a,b]} \left\| X_j - Z_j(I) \right\|_{L^2[a,b]} \|\beta_n(\cdot,s)\|_{L^2[a,b]} \left\| Z_j(I) - \hat{X}_j(I) \right\|_{L^2[a,b]} ds \\
= & \frac{2}{n} \sum_{j=1}^{n} \left\| X_j - Z_j(I) \right\|_{L^2[a,b]} \left\| Z_j(I) - \hat{X}_j(I) \right\|_{L^2[a,b]} \int_c^d \|\beta_n(\cdot,s)\|^2_{L^2[a,b]} ds \\
= & 2 \|\beta_n\|^2_{L^2([a,b]\times[c,d])} \frac{1}{n} \sum_{j=1}^{n} \left\| X_j - Z_j(I) \right\|_{L^2[a,b]} \left\| Z_j(I) - \hat{X}_j(I) \right\|_{L^2[a,b]} \\
\leq & 2 \|\beta_n\|^2_{L^2([a,b]\times[c,d])} \left(\frac{1}{n} \sum_{j=1}^{n} \left\| X_j - Z_j(I) \right\|^2_{L^2[a,b]} \right)^{1/2} \left(\frac{1}{n} \sum_{j=1}^{n} \left\| Z_j(I) - \hat{X}_j(I) \right\|^2_{L^2[a,b]} \right)^{1/2}.
\end{aligned}
$$

El término $\|\beta_n\|^2_{L^2([a,b]\times[c,d])}$ converge por **HfRf1** en probabilidad a $\|\beta_0\|^2_{L^2([a,b]\times[c,d])}$, por lo tanto está acotado en probabilidad. Por **Hf2** y la Ley Fuerte de los Grandes Números el término $\frac{1}{n} \sum_{j=1}^{n} \left\| X_j - Z_j(I) \right\|^2_{L^2[a,b]}$ converge c.s. Y por **Hf1** converge a cero c.s. $\frac{1}{n} \sum_{j=1}^{n} \left\| Z_j(I) - \hat{X}_j(I) \right\|^2_{L^2[a,b]}$. Por consiguiente el producto de estos tres términos converge a cero en probabilidad.

Luego demostramos que $h_n(I) \to_{p.} h(I)$, por consiguiente por el Lema 1, $arg\min_{I \in \mathcal{I}_d} h_n(I) \to_{p.} arg\min_{I \in \mathcal{I}_d} h(I)$, y por tanto, $P(I_n \notin \mathcal{I}_0) \to_{c.s.} 0$ cuando $n \to \infty$.

Bibliografía

Akaike, H. (1974). "A new look at the statistical identification model." *IEEE Trans. Auto. Control.* **19**, 716–723.

Brehency, P. y Huang, J. (2011). "Coordinate descent algorithm for nonconvex penalized regression, with applications to biological feature selection." *Annals of Applied Statistics.* **5**(1), 232–253.

Cadima, J. y Jolliffe, I.T. (2001). "Variable selection and the interpretation of principal subspaces. J. Agricultural." *Biological and Environmental Statistics.* **6**(1), 62–79.

Cardot, H., Ferraty, F. y Sarda, P. (1999). "Functional linear model." *Statistics & Probability Letters.* **45**, 11–22.

Cardot, H., Ferraty, F. y Sarda, P. (2003). "Spline estimators for the functional linear model." *Statistica Sinica.* **13**, 571–591.

Cai, T. T. y Hall, P. (2006). "Prediction in Functional Linear Regression." *The Annals of Statistics.* **34**(5), 2159–2179.

Dauxois, J., Pousse, A. y Romain, Y. (1982). "Asymptotic theory for the principal component analysis of a vector random function: Some applications to statistical inference." *Journal of Multivariate Analysis.* **12**, 136–154.

Devroye, L. (1981). "On the almost everywhere convergence of nonparametric regression function estimates." *Annals of Statistics.* **9**(6), 1310–1319.

Devroye, L. (1982). "Necessary and sufficient conditions for the pointwise convergence of nearest neighbor regression function estimates." *Z. Wahrsch. Verw. Gebiete.* **61**(4), 467–481.

Efron, B., Hastie, T., Johnstone, I. y Tibshirani, R. (2004). "Least Angle Regression." *Annals of Statistics.* **32**(2), 407–451.

Fan, J. y Li, R. (2001). "Variable selection via nonconcave penalized likelihood and its oracle properties." *Journal of American Statistical Association.* **96**(456), 1348–1360.

Ferraty, F. y Romain, Y. (2011). "The Oxford handbook of functional data analysis." *New York, Oxford.*

Fraiman, R., Justel, A. y Svarc, M. (2008). "Selection of Variables for Cluster Analysis and Classification Rules." *Journal of American Statistical Association.* **103**(483), 1294–1303.

Frank, A. y Asuncion, A. (2010). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml], Irvine, CA: University of California, School of Information and Computer Science.

Grenander, U. (1950). "Stochastic process and statistical inference" *Arkiv för Matematik* **1**, 195–277.

Hall, P. y Horowitz, J. L. (2007). "Methodology and convergence rate for functional linear regression." *The Annals of Statistics.* **35**(1), 70–91.

Hastie, T., Tibshirani, R. y Friedman, J. (2001). "The elements of Statistical Learning: Data Mining, Inference and Prediction." *Springer.*

Horváth, L. y Kokoszka, P. (2012). "Inference for functional data with applications." *New York, Springer.*

James, G. M., Wang, J. y Zhu, J. (2009). "Functional linear regression that's interpretable." *The Annals of Statistics.* **37**(5A), 2083–2108.

Jeffers, J. (1967). "Two case studies in the application of principal component." *Journal of the Royal Statistical Society, Series C (Applied Statistics).* **16**(3), 225–236.

Jolliffe, I.T. (2002). "Principal Components Analysis. Second Edition." *Ed. Springer.*

Jolliffe, I., Trendafilov, N. y Uddin, M. (2003). "A modified principal component technique based on the LASSO." *Journal of Computational and Graphical Statistics.* **12**, 531–547.

Li, R. y Gong, G. (1987). "K-nn Nonparametric Estimation Of Regression Functions In the Presence of Irrelevant Variables." *Econometrics Journal.* **00**, 1–12.

Li, Y. y Hsing, T. (2007). "On rates of convergence in functional linear regression." *Journal of Multivariate Analysis.* **98**, 1782–1804.

Lian, H.(2011). " Convergence of functional k-nearest neighbor regression estimate with functional responses." *Electronic Journal of Statistics.* **5**, 31–40.

Maronna, R., Martin, R. D. y Yohai, V. J. (2006). "Robust Statistics: Theory and Methods." *John & Sons Wiley, London.*

McCabe, G. P. (1984). "Principal variables." *Technometrics.* **26**, 137–144.

McCullagh, P. y Nelder, J. (1989).“Generalized Linear Models.” *Chapman and Hall, London.*

Müller, H. G. y Yao, F. (2008). “Functional additive models.” *Journal of American Statistical Association.* **103**, 1534–1544.

Ramsay, J. y Dalzell, C.J. (1991). “Some tools for functional data analysis (with discussion).” *Journal of the Royal Statistical Society. Series B.* **53**, 539–572.

Ramsay, J. y Silverman, B.W. (2005). “Functional Data Analysis(Second Edition).” *New York, Springer.*

Riesz, F. y Nagy, B. (1965). “Lecons d’analyse functionelle.” *Gauthiers-Villars, Paris.*

Rousseauw, J., du Plessis, J., Benade, A., Jordaan, P., Kotze, J., Jooste, P. y Ferreira, J. (1983). “Coronary risk factor screening in three rural communities.” *South African Medical Journal.* **64**, 430–436.

Schwarz, G. (1978). “Estimating the dimension of a model.” *Annals of Statistics.* **6**, 461–464.

Tian, T. S. y James, G. M. (2013). “Interpretable Dimension Reduction for Classifying Functional Data.” *Journal of Computational Statistics and Data Analysis.* **57**(1), 282–296.

Tibshirani, R. (1996). “Regression Shrinkage and Selection via the Lasso.” *Journal of the Royal Statistical Society. Series B* **58**(1), 267–288.

Wherry, R. J. (1931) “A new formula for predicting the shrinkage of the coefficient of multiple correlation.” *Annals of Mathematical Statistics.* **2**, 440–451.

Witten, D. M., Tibshirani, R. y Hastie, T. (2009). “A penalized matrix descomposition , with applications to sparse principal components and canonical correlation analysis.” *Biostatistics.* **10**(3), 515–534.

Yao, F., Müller, H. G. y Wang, J. L. (2005). “Functional linear regression analysis for longitudinal data.” *The Annals of Statistics.* **33**(6), 2873–2903.

Zhang, C. H. (2010). “Nearly unbiased variable selection under minimax concave penalty.” *Annals of Statistics.* **38**, 894–942.

Zou, H. y Hastie, T. (2005). “‘Regularizations and variable selection via the elastic net.” *Journal of the Royal Statistical Society. Series B.* **67**, 301–320.

Zou, H., Hastie, T. y Tibshirani, R. (2006). “Sparce principal component analysis.” *Journal of Computational and Graphical Statistics.* **15**, 265–286.