UNIVERSIDAD DE BUENOS AIRES

Facultad de Ciencias Exactas y Naturales

Maestría en Estadística Matemática

Una propuesta de estimación y selección simultánea de variables en modelos parcialmente lineales aditivos

Tesis presentada para optar por el título de Magíster de la Universidad de Buenos Aires en Estadística Matemática

Nicolás Murrone

Directora: Dra. Alejandra M. Martínez

20 de noviembre de 2025

Resumen

Para estudiar la relación entre una variable respuesta y un conjunto de variables explicativas, los modelos de regresión parcialmente lineales aditivos, que se encuentran dentro de los modelos de regresión semiparamétricos, constituyen una herramienta útil, ya que permiten incluir componentes no paramétricas aditivas junto con una componente paramétrica lineal.

En la práctica, es usual recolectar un gran número de covariables, aunque no todas ellas resulten relevantes para el modelo. En este contexto, los modelos esparsos suponen que la cantidad de predictores relevantes es menor que el número de covariables medidas, lo que lleva a modelos más fáciles de interpretar que los modelos densos. Por esta razón, la selección de variables resulta crucial, dado que la inclusión de covariables sin impacto en la variable respuesta suele reducir la capacidad predictiva del modelo.

En este trabajo nos ocuparemos de modelos parcialmente lineales aditivos bajo supuestos de esparsidad tanto en la componente lineal como en la componente no paramétrica. Para ello, se propone un procedimiento de estimación y selección simultánea de variables que utiliza funciones $B-$splines para aproximar las componentes aditivas, junto con un procedimiento de regularización del tipo *adaptativo*. Además, se proponen criterios del tipo BIC para seleccionar de forma automática tanto los parámetros de suavidad relacionados con la aproximación por splines como los parámetros de penalidad que controlan la esparsidad del modelo. La propuesta se evalúa mediante estudios de simulación bajo distintos escenarios y se ilustra su aplicación a través del análisis de un conjunto de datos reales.

Abstract

An estimation and simultaneous variable selection approach for partially linear additive models

To study the relationship between a response variable and a set of explanatory variables, partially linear additive regression models, which belong to the class of semiparametric regression models, provide a useful tool, as they allow the inclusion of additive nonparametric components along with a linear parametric component.

In practice, it is common to collect a large number of covariates, although not all of them may be relevant to the model. In this context, sparse models assume that the number of relevant predictors is smaller than the total number of measured covariates, leading to models that are easier to interpret than dense ones. For this reason, variable selection plays a crucial role, since including covariates that have no impact on the response variable often reduces the predictive performance of the model.

In this work, we focus on partially linear additive models under sparsity assumptions for both the linear and nonparametric components. To this end, we propose an estimation and simultaneous variable selection approach that uses $B-$splines to approximate the additive components, combined with an adaptive regularization procedure. Additionally, we propose BIC-type criteria for automatically selecting both the smoothing parameters related to the spline approximation and the penalty parameters that control model sparsity. The proposed methodology is evaluated through simulation studies under various scenarios and illustrated with the analysis of a real dataset.

Agradecimientos

A mi mamá y mi papá por apoyarme y guiarme en cada uno de los pasos que voy dando en esta vida. No puedo estar más agradecido por su apoyo permanente. Esta tesis está dedicada a ellos y al esfuerzo que han hecho para que sea la persona que hoy soy. Gracias por el amor que me dan cada día.

A mi hermano Mateo: cada día con sus mensajes me hace los días más felices. Agradecido por los eventos deportivos que hemos compartido, y todos los que compartiremos. Su manera de ver la vida, y sus reflexiones siempre me hacen ver las cosas desde otra perspectiva.

A Natalia por el apoyo incondicional en esta última etapa de mis estudios, pero sobretodo por elegir compartir la vida conmigo y hacer reales mis proyectos personales.

A la memoria de mi abuela Cata y mi abuelo Antonio. Puedo imaginarlos sonrientes y felices de saber que he logrado esto. Gracias por el amor que me han brindado y por esa ternura que no he vuelto a ver en nadie.

No es un agradecimiento, pero sí un saludo a Juanca y su pronta partida de este mundo. Siempre recordaremos tu emoción al probar las aceitunas como "las que hacía mi abuela".

A Alejandra, mi directora de tesis, por su gran empeño, por sus ideas, por su paciencia, sus correcciones tan precisas, por guiarme en esta última etapa de mis estudios. Gracias por el tiempo dedicado a este trabajo que hoy llega a su fin gracias a ella.

A Jazmín, mi compañera desde el principio de la especialización, hasta el final de la maestría. Gracias a ella por sus increíbles redacciones en cada uno de los trabajos de cada una de las materias. Gracias a ella por tantas horas de Meet estudiando y riendo en uno de los años más tristes que he vivido.

A mis amigos de toda la vida por siempre estar, por sacar carcajadas como nadie más lo hace. Son uno de mis cables a tierra.

Índice general

1. Introducción **1**

2. Estimación en modelos PLAM **3**

2.1. Caso particular: modelos lineales 4

2.2. Caso particular: modelos aditivos 6

2.3. Estimadores basados en $B-$splines 8

2.3.1. Splines y $B-$splines 8

2.3.2. Propuesta de Boente y Martínez (2023) 11

3. Selección de variables **14**

3.1. Método: "mejor subconjunto" (*best subset*) 15

3.2. Métodos de inclusión y exclusión de variables 16

3.2.1. Método: "hacia adelante" (selección *forward*) 17

3.2.2. Método: "hacia atrás" (eliminación *backward*) 17

3.2.3. Método: "paso a paso" (*stepwise*) 18

3.3. Métodos de regularización 20

3.3.1. Estimador Ridge 21

3.3.2. Estimador LASSO 22

3.3.3. Estimador ADALASSO 22

3.3.4. Estimadores Bridge y Elastic Net 23

3.3.5. Estimador con penalidad MCP 24

3.3.6. Estimador con penalidad SCAD 25
3.4. Selección de variables en modelos semiparamétricos esparsos 28

4. La propuesta **31**
4.1. Estimación y selección simultánea de variables en modelos PLAM . 31
4.2. Selección del parámetro de penalidad 33
4.3. Selección del parámetro de suavizado 34
4.4. Algoritmo general 35

5. Estudio de simulación **40**
5.1. Escenarios 41
5.2. Medidas de selección de variables 46
5.3. Medidas de estimación 47
5.4. Resultados 48
5.4.1. Sobre la calidad de la selección de variables 49
5.4.2. Sobre la calidad de la estimación 52
5.4.3. Sobre la selección de los parámetros de penalización 79
5.4.4. Sobre la selección de los parámetros de suavizado 84

6. Análisis de datos reales **89**
6.1. Incorporación de covariables 96

7. Conclusiones **102**

Capítulo 1

Introducción

Los modelos de regresión parcialmente lineales aditivos (PLAM) forman una clase de modelos semiparamétricos que combinan componentes lineales y no paramétricas para modelar la relación entre una variable respuesta y un conjunto de variables explicativas. Estos modelos proveen una herramienta poderosa para abordar situaciones donde algunas covariables se relacionan de manera lineal con la respuesta, mientras que otras tienen una relación no paramétrica. Esta flexibilidad los hace una alternativa interesante frente a los modelos de regresión lineales convencionales, al tiempo que permiten una interpretación más clara que los modelos no paramétricos puros (ver Hastie y Tibshirani, 1990).

En la práctica, se suele recolectar un gran número de covariables, aunque no todas ellas resultan ser relevantes para el modelo. En este contexto, los modelos esparsos ofrecen una solución eficiente, ya que suponen que sólo un pequeño subconjunto de las covariables tiene un impacto significativo en la variable respuesta. Este enfoque permite construir modelos más simples y fácilmente interpretables, evitando la sobrecarga de variables irrelevantes que podrían reducir la capacidad predictiva del modelo. La selección de variables, por tanto, juega un rol fundamental en la mejora de los modelos estadísticos. Incluir covariables que no aportan valor a la respuesta puede dificultar la interpretación del modelo y afectar su desempeño predictivo.

El uso de métodos de penalización, que utilizan funciones de penalidad como la L_1, que da origen al estimador LASSO, o las penalidades SCAD o MCP, entre otras, es una estrategia común para realizar esta selección de manera automática, facilitando la identificación de las variables más relevantes. Estos métodos no sólo ayudan a mejorar la interpretación del modelo, sino que también tienen la ventaja de tener bajos costos computacionales, diferenciándose así de otros procedimientos de selección de variables como, por ejemplo, *best subset* o los métodos

del tipo *stepwise*. Una discusión detallada sobre estas diferencias metodológicas puede encontrarse en Hastie (2017).

En este trabajo abordamos los modelos parcialmente lineales aditivos bajo suposiciones de esparsidad tanto para las covariables incluidas en la componente lineal como en las componentes no paramétricas. Proponemos un enfoque de estimación combinado con un procedimiento de regularización adaptativo que utiliza $B-$splines para aproximar las componentes aditivas y emplea un criterio de penalización SCAD para seleccionar las variables relevantes. Además, se propone el uso de criterios tipo BIC para seleccionar automáticamente los parámetros de suavidad y de penalización. Finalmente, a través de un estudio de simulación y un análisis de datos reales, ilustramos la efectividad de la metodología propuesta y sus ventajas en términos de interpretación y capacidad predictiva.

El presente trabajo de tesis se estructura de la siguiente manera. En el Capítulo 2, se realiza un recorrido por los métodos de estimación tradicionales aplicados a los modelos parcialmente lineales aditivos (y a sus casos particulares de modelos lineales y modelos aditivos), destacando tanto las ventajas como las limitaciones de los enfoques más utilizados. En el Capítulo 3, se repasan las principales técnicas clásicas de selección de variables, como el método de "mejor subconjunto" (*best subset*) y los del tipo *stepwise selection*, y se ofrece una visión general de los métodos de penalización más relevantes, incluyendo LASSO, Elastic Net, MCP y SCAD. En el Capítulo 4, se presenta una propuesta innovadora de estimación para estos modelos, la cual busca mejorar la precisión y la parsimonia en la selección de variables mediante el uso de técnicas avanzadas de regularización y estimación simultánea en la componente lineal y en la componente aditiva. Posteriormente, en el Capítulo 5, se lleva a cabo un estudio de simulación para evaluar el rendimiento de los métodos propuestos en distintos escenarios. Finalmente, en el Capítulo 6, se aplica nuestra propuesta de estimación a un conjunto de datos reales de *plasma*, ilustrando su aplicabilidad y mostrando los resultados obtenidos en un contexto práctico.

Capítulo 2

Estimación en modelos PLAM

Los modelos parcialmente lineales aditivos (PLAM, por sus siglas en inglés), que contienen una componente lineal y una componente no paramétrica aditiva, son más flexibles que los modelos lineales, de fácil interpretación y resuelven el problema de la "maldición de la dimensión" que suelen poseer los modelos completamente no paramétricos.

En estos modelos, la variable respuesta Y es modelada usando un conjunto de covariables, el cual es dividido en dos subconjuntos: el de variables que ingresan al modelo de forma lineal, y el de las variables que son incluidas al modelo a través de funciones suaves desconocidas. Más precisamente, estos modelos asumen que se tienen $(Y_i, \mathbf{Z}_i^{\mathrm{T}}, \mathbf{X}_i^{\mathrm{T}})^{\mathrm{T}} \in \mathbb{R}^{1+q+p}$, $1 \leq i \leq n$, vectores aleatorios independientes e idénticamente distribuidos con la misma distribución que $(Y, \mathbf{Z}^{\mathrm{T}}, \mathbf{X}^{\mathrm{T}})^{\mathrm{T}}$ tal que

$$Y = m(\mathbf{Z}^{\mathrm{T}}, \mathbf{X}^{\mathrm{T}}) + u = \mu + \boldsymbol{\beta}^{\mathrm{T}}\mathbf{Z} + \sum_{j=1}^{p} \eta_j(X_j) + u \tag{2.1}$$

donde los errores u se suponen independientes del vector de covariables $(\mathbf{Z}^{\mathrm{T}}, \mathbf{X}^{\mathrm{T}})^{\mathrm{T}}$. En general, se los considera con media 0 y varianza σ^2. Además la constante $\mu \in \mathbb{R}$, el vector $\boldsymbol{\beta} \in \mathbb{R}^q$ y las funciones univariadas $\eta_j : \mathbb{R} \to \mathbb{R}$, $1 \leq j \leq p$, son los objetos a estimar. Sin embargo, este modelo así como está definido no es identificable. Para asegurar identificabilidad de las componentes aditivas η_j y de μ, se asume que $\int \eta_j(x)\, dx = 0$ y que $\eta_j : \mathcal{I}_j \to \mathbb{R}$ y entonces $\int_{\mathcal{I}_j} \eta_j(x)\, dx = 0$, para $1 \leq j \leq p$. Más aún, sin pérdida de generalidad, asumiremos que $\mathcal{I}_j = [0, 1]$ para $1 \leq j \leq p$.

El modelo de regresión parcialmente lineal aditivo (2.1) incluye como casos particulares al modelo de regresión lineal cuando $\eta_j \equiv 0$ para $1 \leq j \leq p$ (ver Sección 2.1), al modelo parcialmente lineal cuando $p = 1$ y al modelo aditivo cuando el parámetro de regresión es cero, es decir, cuando $q = 0$ (ver Sección 2.2).

Cabe destacar que los modelos PLAM permiten incluir en la componente lineal covariables que sean discretas, dummies o no acotadas. Además, como se menciona en Opsomer y Ruppert (1999), proporcionan un modelo parsimonioso que puede interpretarse fácilmente y que son adecuados cuando se está bastante seguro de la relación lineal entre la respuesta y algún subconjunto de covariables, pero no de la relación con las otras.

En lo que respecta a la estimación, Guo et al. (2013) propusieron usar una base de splines polinomiales para aproximar las componentes no paramétricas aditivas del modelo (2.1) y luego un procedimiento basado en regresión por cuantiles. Siguiendo esta línea de investigación, Boente y Martínez (2023) propusieron una familia de estimadores robustos para modelos PLAM combinando $B-$splines con estimadores robustos en regresión lineal. Con esta aproximación en la parte aditiva, cada componente no paramétrica es representada por una combinación lineal de la base de funciones de splines.

El presente capítulo se organiza de la siguiente manera. En la Sección 2.1 se hace un breve repaso de dos estimadores clásicos para modelos lineales: el estimador de mínimos cuadrados y el estimador Ridge. En la Sección 2.2 se mencionan diferentes métodos de estimación en modelos aditivos. En la Sección 2.3 se menciona el método de estimación en modelos PLAM basado en $B-$splines propuesto en Boente y Martínez (2023) y que nos servirá como base para la propuesta de estimación y selección simultánea de variables desarrollada en la presente tesis.

2.1. Caso particular: modelos lineales

Los modelos lineales clásicos son un caso particular de los modelos PLAM al tomar $p = 0$ en (2.1). Suponen que se tienen $(Y_i, \mathbf{Z}_i^{\mathrm{T}})^{\mathrm{T}} \in \mathbb{R}^{1+q}$, $1 \leq i \leq n$, vectores aleatorios independientes e idénticamente distribuidos con la misma distribución que $(Y, \mathbf{Z}^{\mathrm{T}})^{\mathrm{T}}$ tal que

$$Y = \mu + \boldsymbol{\beta}^{\mathrm{T}}\mathbf{Z} + u,$$

donde la constante $\mu \in \mathbb{R}$ y el vector $\boldsymbol{\beta} \in \mathbb{R}^q$ son los objetos a estimar. Los errores u se suponen independientes del vector de covariables $\mathbf{Z}$ y, al igual que en los modelos PLAM, con media 0 y varianza σ^2.

Para estimar los valores de μ y $\boldsymbol{\beta}$, el método usual es el *estimador de mínimos cuadrados*. Para valores genéricos de $a \in \mathbb{R}$ y $b \in \mathbb{R}^q$, definimos los residuos $r_i(a, \mathbf{b})$ como $r_i(a, \mathbf{b}) = Y_i - a - \mathbf{b}^{\mathrm{T}}\mathbf{Z}_i$. Luego, el estimador de mínimos cuadrados se define minimizando la suma de los cuadrados de los residuos, es decir, resolviendo el

problema de optimización

$$(\widehat{\mu}, \widehat{\boldsymbol{\beta}}) = \underset{a\in\mathbb{R},\mathbf{b}\in\mathbb{R}^q}{\operatorname{argmin}} \sum_{i=1}^{n} r_i^2(a, \mathbf{b}).$$

Bajo los *supuestos clásicos* de regresión (linealidad, independencia de los errores, homocedasticidad), el estimador de mínimos cuadrados es insesgado y consistente, lo que garantiza que, a medida que el tamaño de la muestra n crece, el estimador se aproxima al verdadero valor de los parámetros del modelo. Gujarati (2009) explica que el estimador de mínimos cuadrados tiene propiedades deseables, como la eficiencia de los estimadores cuando los errores son homocedásticos y no están correlacionados. Sin embargo, en la práctica, es común que estos supuestos no se cumplan perfectamente. De hecho, en presencia de multicolinealidad, donde las variables están altamente correlacionadas entre sí, los estimadores obtenidos por mínimos cuadrados pueden volverse ineficientes. Como explica Wooldridge (2010), la multicolinealidad no impide que el estimador sea insesgado, pero sí dificulta la interpretación de los coeficientes y aumenta la varianza de los estimadores. Kutner et al. (2004) discuten cómo el aumento del número de variables puede conducir a problemas de multicolinealidad, lo que provoca que los coeficientes estimados sean imprecisos y difíciles de interpretar. Adicionalmente, cuando se tiene un número elevado de variables, el modelo puede sufrir de sobreajuste, es decir, el modelo ajusta demasiado los datos de entrenamiento, capturando ruido en lugar de patrones reales, lo que limita la capacidad de predicción.

Cuando hay muchas variables correlacionadas en un modelo de regresión lineal, sus coeficientes pueden quedar mal determinados y exhibir una alta varianza. Un coeficiente positivo tremendamente grande en una variable puede ser cancelado por un coeficiente negativo igualmente grande en otra variable correlacionada. Para abordar estos problemas, la regresión Ridge propuesta por Hoerl y Kennard (1970) introduce una penalización sobre la magnitud de los coeficientes, lo que ayuda a estabilizar las estimaciones y a reducir el sobreajuste. Este método se obtiene al modificar el problema de estimación clásico, añadiendo un término de penalización basado en la norma L_2 de los coeficientes del vector $\mathbf{b}$. Además, cabe destacar que, en la práctica, la constante μ se estima de forma preliminar, tal como se menciona en Boente y Martínez (2023), y, por lo tanto, sin pérdida de generalidad se puede considerar $\mu = 0$. En caso contrario, las respuestas Y_i se redefinen como $Y_i - \widehat{\mu}$ y se procede de forma análoga. En definitiva, el estimador Ridge se obtiene resolviendo el siguiente problema de optimización:

$$\widehat{\boldsymbol{\beta}}^{\text{RIDGE}} = \underset{\mathbf{b}\in\mathbb{R}^q}{\operatorname{argmin}} \sum_{i=1}^{n} \left(Y_i - \sum_{j=1}^{q} b_j Z_{ij} \right)^2 + \lambda \sum_{j=1}^{q} b_j^2 \,, \tag{2.2}$$

con $\mathbf{b} = (b_1, \ldots, b_q)^{\mathrm{T}} \in \mathbb{R}^q$ y donde $\lambda \geq 0$ es un parámetro de regularización que controla la magnitud de la penalización: a medida que λ se hace más grande, el encogimiento es cada vez mayor, reduciendo los coeficientes hacia cero; en cambio, a medida que λ se acerca cada vez más a cero, predomina el primero de los términos de la función objetivo, asemejándose este procedimiento al de mínimos cuadrados y, por lo tanto, no reduce el tamaño de los coeficientes.

La regresión Ridge presenta varias ventajas, entre ellas su estabilidad numérica en presencia de multicolinealidad y su capacidad para mejorar el desempeño predictivo del modelo al aplicarlo a nuevos datos. Sin embargo, a pesar de estas ventajas, una de sus principales limitaciones es que no realiza una selección explícita de variables, lo que significa que todas las variables siguen contribuyendo al modelo, aunque sus coeficientes estén más reducidos. Este aspecto puede ser una desventaja cuando se desea identificar un subconjunto de variables más relevantes para el modelo. Además, como señalan James et al. (2013), en muchos casos, el uso de todas las variables disponibles puede resultar en modelos sobreajustados, donde los estimadores de los coeficientes no son confiables y las predicciones del modelo tienen un alto error en datos no observados. Aunque la regresión Ridge mitiga parte del sobreajuste, sigue sin abordar el problema de las variables irrelevantes, lo que puede afectar la interpretabilidad y la capacidad predictiva del modelo. Estas limitaciones del estimador de mínimos cuadrados, junto con las de Ridge, resaltan la importancia de los métodos de selección de variables, los cuales ayudan a identificar las variables más relevantes y a eliminar las que no aportan valor al modelo. En la Sección 3.3 exploraremos otras técnicas de penalización que sí efectúan una selección explícita de variables, como puede ser la utilización de la penalidad L_1 que da origen al estimador LASSO, y la utilización de otras funciones de penalidad como la Elastic Net, SCAD o MCP.

2.2. Caso particular: modelos aditivos

Los modelos aditivos permiten modelar a la variable respuesta Y como la suma de p funciones suaves evaluadas cada una de ellas en una única componente del vector de covariables $\mathbf{X} = (X_1, \ldots, X_p)^{\mathrm{T}}$. La ventaja de los modelos aditivos respecto del modelo puramente no paramétrico es que permiten resolver la llamada "maldición de la dimensión", la cual es causada por el hecho de que el número esperado de observaciones a nivel local disminuye exponencialmente en función de la dimensión p de las covariables. Esto resulta en una pobre tasa de convergencia de los estimadores que depende exponencialmente de la dimensión y del grado de suavidad de la función de regresión. Más precisamente, si $(Y, \mathbf{X}^{\mathrm{T}})$ es un vector aleatorio, $Y \in \mathbb{R}$ es la variable dependiente y $\mathbf{X} \in \mathbb{R}^p$ es el vector de covariables

explicativas, el modelo de regresión puramente no paramétrico se define como

$$Y = \eta(\mathbf{X}) + u,$$

donde el error u es independiente de $\mathbf{X}$, centrado en cero, y $\eta : \mathbb{R}^p \to \mathbb{R}$ es la función a ser estimada. En Stone (1985) se mostró que la tasa óptima de convergencia para estimar η es $n^{\ell/(2\ell+p)}$ donde ℓ es el grado de suavidad de η. Por lo tanto, altos valores de p llevan a una tasa de convergencia más lenta. Este hecho está asociado al fenómeno de entornos cada vez más esparsos al aumentar la dimensión.

Para resolver el problema de la "maldición de la dimensión" de los modelos completamente no paramétricos, Stone (1985) y Hastie y Tibshirani (1986) consideraron modelos aditivos, los cuales, además de resolver la "maldición de la dimensión", generalizan a los modelos lineales y son fácilmente interpretables. Estos modelos suponen que $\eta(x) = \mu + \sum_{j=1}^{p} \eta_j(x_j)$, donde $\mu \in \mathbb{R}$ es el parámetro de centralidad, y las componentes aditivas $\eta_j : \mathcal{I}_j \subset \mathbb{R} \to \mathbb{R}$ satisfacen alguna condición de identificabilidad como $E\eta_j(X_j) = 0$ o bien $\int_{\mathcal{I}_j} \eta_j(x)\, dx = 0$, para $1 \leq j \leq p$. En definitiva, en los modelos aditivos, la relación entre la variable respuesta Y y el vector de covariables $\mathbf{X}$ viene dado de la siguiente manera

$$Y = \mu + \sum_{j=1}^{p} \eta_j(X_j).$$

En la literatura se han propuesto varios procedimientos para estimar modelos aditivos. Uno de los más utilizados es el método iterativo denominado backfitting, formalmente introducido por Buja et al. (1989) y Hastie y Tibshirani (1990). La convergencia y el análisis formal de este algoritmo han sido abordados por Opsomer y Ruppert (1997). Si bien es un procedimiento que, usualmente, converge rápidamente, su naturaleza iterativa dificulta el análisis de sus propiedades estadísticas. Por otro lado, el procedimiento denominado "integración marginal" propuesto por Tjøstheim y Auestad (1994) y Linton y Nielsen (1995) admite una fórmula cerrada para el estimador pero resulta en un procedimiento más costoso computacionalmente.

Una de las ventajas de los modelos aditivos es que permiten interpretar independientemente los efectos de cada covariable sobre la función de regresión η, tal como sucede con los modelos de regresión lineal. Además, como fue mostrado por Stone (1985), para estos modelos aditivos, la tasa de convergencia óptima para estimar η es una tasa de convergencia unidimensional $n^{\ell/(2\ell+1)}$, reduciendo la dimensionalidad del problema a través de un modelo aditivo. Vale la pena resaltar que los modelos aditivos combinan la flexibilidad de los modelos no paramétricos con la simple interpretación del clásico modelo lineal. Sin embargo, una de las

mayores desventajas de estos modelos es que las variables involucradas deben ser continuas y, por lo tanto, variables del tipo dummies o discretas no pueden formar parte del vector $\mathbf{X}$. Como ya se mencionó, el modelo aditivo es un caso particular del modelo PLAM tomando $q = 0$ en la expresión (2.1). En consecuencia, este modelo no permite incluir covariables de forma lineal, lo cual puede ser una limitación si se cuenta con información que indique que algunas covariables tienen una relación lineal con la variable respuesta, o si se trata de variables no continuas que deben incorporarse de manera paramétrica.

2.3. Estimadores basados en $B-$splines

A continuación, se describe la propuesta de estimación desarrollada en Boente y Martínez (2023) para modelos PLAM, la cual servirá como base para el desarrollo del enfoque metodológico de estimación y selección simultánea de variables propuesto en esta tesis. En términos generales, dicha propuesta consiste en aproximar las funciones aditivas mediante combinaciones lineales de funciones $B-$splines, con el objetivo de representar el modelo PLAM como un modelo lineal y, de este modo, poder aplicar herramientas clásicas de estimación, como la estimación por mínimos cuadrados.

2.3.1. Splines y $B-$splines

Para aproximar cada función aditiva η_j de un modelo parcialmente lineal aditivo, se utilizará a una base de funciones, elegida individualmente para cada componente no lineal. Esta construcción por separado permite capturar adecuadamente las particularidades de cada η_j, como su grado de suavidad o forma general. Una opción ampliamente utilizada para este fin consiste en emplear funciones *splines*, que permiten representar funciones suaves a partir de combinaciones lineales de elementos base.

Un *spline* de grado D_j definido sobre un intervalo $[a, b]$ con $N_{n,j}$ nodos internos establecidos para la j-ésima función aditiva, es una función continua a tramos que, en cada subintervalo determinado por dos nodos consecutivos, coincide con un polinomio de grado a lo sumo D_j, y cuya derivada de orden hasta $D_j - 1$ es continua en todo el dominio. Es decir, un spline es una función que combina polinomios locales, unidos en los nodos de forma suficientemente suave. Además, se define el *orden* de un spline, denotado aquí por ℓ_j, como $\ell_j = D_j + 1$.

Dentro de las funciones splines, los $B-$splines representan una clase particular

que se caracteriza por tener soporte compacto, ser no negativos y formar una base para el espacio de funciones splines de determinado grado y nodos. Cada función $B-$spline está definida localmente sobre un número reducido de subintervalos y es polinomial en cada uno de ellos. Los $B-$splines se construyen a partir de un conjunto ordenado de nodos y el orden de la función. Existe un *teorema fundamental*, propuesto por Schoenberg (1946), que establece que cualquier spline de grado D_j definido sobre un conjunto de nodos dado, puede expresarse de manera única como una combinación lineal de funciones $B-$spline de ese mismo grado asociadas a dichos nodos. En particular, este resultado demuestra que los $B-$splines forman una base del espacio de funciones splines correspondiente. Además, se sabe que la cantidad de funciones de la base necesarias, denotada como k_j, viene dada por $k_j = k_{n,j} = N_{n,j} + D_j + 1$ o, equivalentemente, $k_j = k_{n,j} = N_{n,j} + \ell_j$. Esta cantidad determina cuántas funciones $B-$spline se requieren para generar, mediante combinaciones lineales, cualquier spline de grado D_j y con los nodos previamente especificados.

Esta propiedad hace que los $B-$splines sean una herramienta fundamental para la representación y manipulación de splines en aplicaciones teóricas y prácticas. En particular, como el número de funciones de la base k_j depende tanto del número de nodos internos $N_{n,j}$ como del grado D_j de los polinomios locales, al momento de construir la base, la elección del número de nodos internos y del grado debe hacerse cuidadosamente para garantizar que el espacio generado por las funciones $B-$splines resulte suficientemente flexible para representar bien las funciones deseadas. En definitiva, la construcción de bases de $B-$splines involucra dos decisiones principales: la cantidad y disposición de los nodos internos, y el grado de las funciones polinomiales locales. Una vez fijados estos elementos, se determina de manera directa la cantidad de funciones de base necesarias, que permiten representar cualquier spline en dicho espacio como una combinación lineal de ellas.

La Figura 2.1 muestra las bases de $B-$splines para cuatro grados $D_j = 0, 1, 2, 3$ donde no hay ningún nodo interno. A medida que se aumenta el grado de los $B-$splines, se incrementa la suavidad de las funciones que conforman la base, lo que refleja un mayor nivel de continuidad en las derivadas de orden superior. Es interesante observar que cuando no hay nodos internos (es decir, $N_j = 0$), las funciones $B-$splines se reducen a funciones polinomiales simples de grado 0, 1, 2, etc., dependiendo del grado especificado. Cuando se introduce un nodo interno, por ejemplo en 0.5, como se muestra en la Figura 2.2, las funciones $B-$splines se adaptan a esta nueva partición del dominio. Cada función base sigue teniendo soporte local, pero su forma se vuelve más versátil, con tramos definidos en más subintervalos. Esto permite una mayor flexibilidad para aproximar funciones con cambios de curvatura más bruscos.

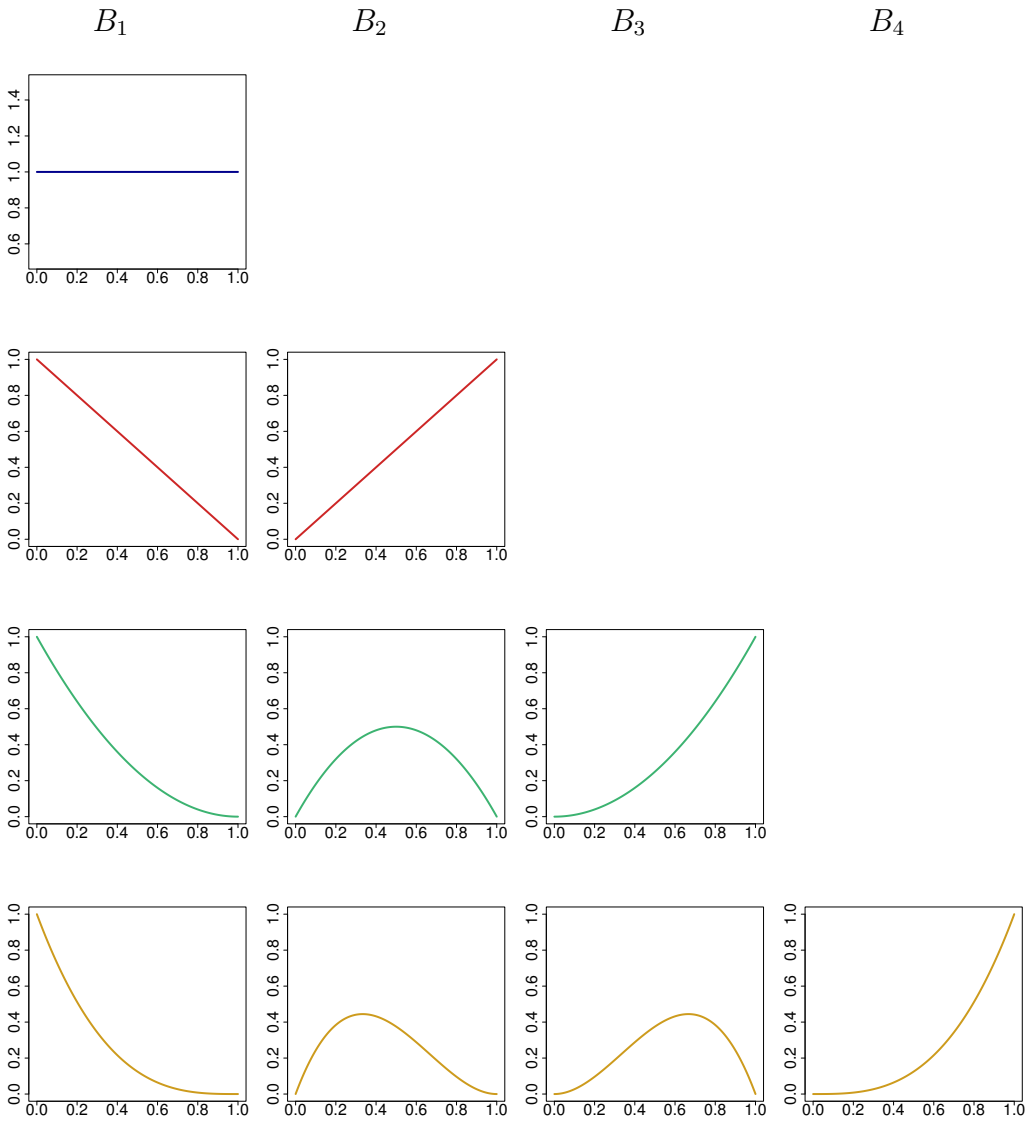

Figura 2.1: Funciones base $B-$spline generadas con $N_j = 0$ nodos internos para grados: $D_j = 0$ en azul, $D_j = 1$ en rojo, $D_j = 2$ en verde y $D_j = 3$ en amarillo. Cada fila muestra las funciones de la base correspondientes a un valor de D_j, identificadas como B_s, con s entre 1 y 4 según corresponda.

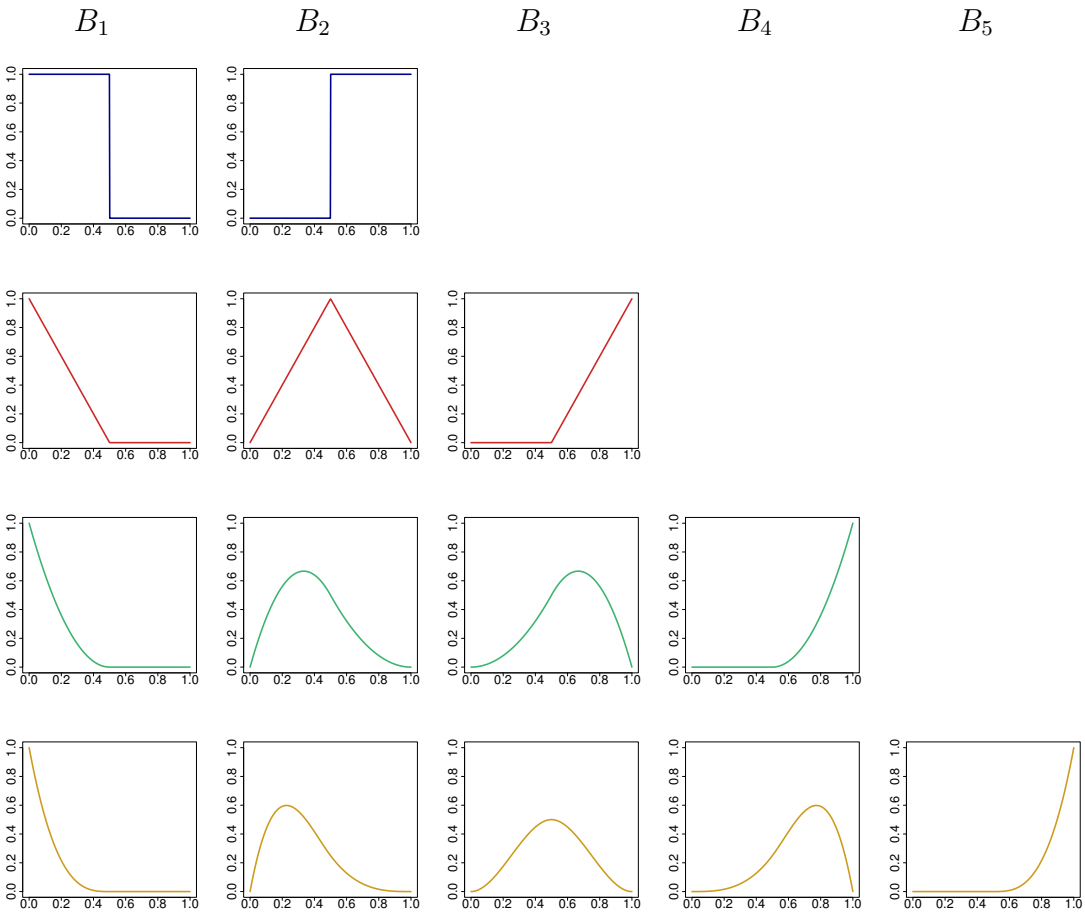

Figura 2.2: Funciones base $B-$spline generadas con un único nodo interno ($N_j = 1$) ubicado en 0.5, para grados: $D_j = 0$ en azul, $D_j = 1$ en rojo, $D_j = 2$ en verde y $D_j = 3$ en amarillo. Cada fila muestra las funciones de la base correspondientes a un valor de D_j, identificadas como B_s, con s entre 1 y 5 según corresponda.

2.3.2. Propuesta de Boente y Martínez (2023)

En lo que sigue describiremos la propuesta desarrollada en Boente y Martínez (2023) que permite obtener los estimadores $\widehat{\mu}$ de μ, $\widehat{\boldsymbol{\beta}}$ de $\boldsymbol{\beta} \in \mathbb{R}^q$ y $\widehat{\eta}_j$ de cada función aditiva η_j, para $j = 1, \ldots, p$.

Para cada $j = 1, \ldots, p$, se fija el orden de spline ℓ_j y el número de nodos internos $N_{n,j}$ y su ubicación, y se obtiene una base de $B-$splines $\{\widetilde{B}_s^{(j)} : 1 \leq s \leq k_j\}$, con $k_j = k_{n,j} = N_{n,j} + \ell_j$ que representa la dimensión. Luego, se define la base centrada de $B-$splines como $\{B_s^{(j)} : 1 \leq s \leq k_j\}$ donde $B_s^{(j)} = \widetilde{B}_s^{(j)} - \int_{\mathcal{I}_j} \widetilde{B}_s^{(j)}(x)\, dx$, es

decir, donde las funciones base son centradas con la integral. De esta forma, esta base que será utilizada para aproximar η_j en (2.1) de la siguiente manera: cada η_j se aproxima como $\sum_{s=1}^{k_j-1} c_s^{(j)} B_s^{(j)}(t)$, es decir, como combinación lineal de los primeros $k_j - 1$ elementos de la base de $B-$splines centrados. Luego, el modelo (2.1) queda aproximado como

$$Y = \mu + \boldsymbol{\beta}^{\mathrm{T}}\mathbf{Z} + \sum_{j=1}^{p}\sum_{s=1}^{k_j-1} c_s^{(j)} B_s^{(j)}(X_j) + u\,.$$

Es interesante observar que, tal como se menciona en Boente y Martínez (2024), en contraposición con los métodos basados en núcleos, los métodos basados en splines proveen estimadores simultáneos del parámetro $\boldsymbol{\beta}$ y de las componentes no paramétricas y no es necesario recurrir a estimadores en varios pasos.

Observemos que para definir nuestros estimadores, es necesario recurrir a una base de $B-$splines centrados. Supongamos que $\{\widetilde{B}_s^{(j)} : 1 \leq s \leq k_j\}$ es una base de $B-$splines. Sea $\boldsymbol{\gamma}^{(j)} \in \mathbb{R}^{k_j}$, con $\boldsymbol{\gamma}^{(j)} = (\gamma_1^{(j)}, \ldots, \gamma_{k_j}^{(j)})^{\mathrm{T}}$, el vector de coeficientes de la combinación lineal y sea $\sum_{s=1}^{k_j} \gamma_s^{(j)} \widetilde{B}_s^{(j)}(t)$ un spline que se utilizará para aproximar η_j. Dado que una de las hipótesis para tener identificabilidad de las η_j es que $\int_{\mathcal{I}_j} \eta_j(x)\,dx = 0$, los coeficientes del vector $\boldsymbol{\gamma}^{(j)}$ deben satisfacer la restricción de que $\sum_{s=1}^{k_j} \gamma_s^{(j)} \int_{\mathcal{I}_j} \widetilde{B}_s^{(j)}(x)dx = 0$. De esta forma, como hacen Guo et al. (2013) y con el fin de resolver un problema de optimización sin restricciones, centramos la base de funciones. Es decir, definimos $B_s^{(j)}(x) = \widetilde{B}_s^{(j)}(x) - \int_{\mathcal{I}_j} \widetilde{B}_s^{(j)}(x)\,dx$, con lo cual la aproximación para η_j queda $\sum_{s=1}^{k_j} \gamma_s^{(j)} B_s^{(j)}(t)$. De esta manera, para valores de $a \in \mathbb{R}$, $\mathbf{b} \in \mathbb{R}^q$ y $\boldsymbol{\gamma}^{(j)} \in \mathbb{R}^{k_j}$, el modelo (2.1) quedaría aproximado como

$$Y = a + \mathbf{b}^{\mathrm{T}}\mathbf{Z} + \sum_{i=1}^{p}\sum_{s=1}^{k_j} \gamma_s^{(j)} B_s^{(j)}(X_j) + u$$

y entonces el estimador de mínimos cuadrados se obtendría minimizando

$$\sum_{i=1}^{n}\left(Y_i - a - \mathbf{b}^{\mathrm{T}}\mathbf{Z}_i - \sum_{i=1}^{p}\sum_{s=1}^{k_j} \gamma_s^{(j)} B_s^{(j)}(X_{ij})\right)^2.$$

Sin embargo, la matriz de diseño para este problema está mal condicionada, incluso cuando $p = 1$, debido a la constante a. Teniendo en cuenta que $\sum_{s=1}^{k_j} \widetilde{B}_s^{(j)}(x) = 1$, para cualquier $x \in \mathcal{I}_j$, se obtiene que $\sum_{s=1}^{k_j} B_s^{(j)}(x) = 0$. Luego, podemos reescribir la aproximación como

$$\sum_{s=1}^{k_j} \gamma_s^{(j)} B_s^{(j)}(x) = \sum_{s=1}^{k_j-1} \left(\gamma_s^{(j)} - \gamma_{k_j}^{(j)}\right) B_s^{(j)}(x).$$

De esta forma, si definimos $\mathbf{c}^{(j)} = (c_1^{(j)}, \ldots, c_{k_j-1}^{(j)})^{\mathrm{T}} \in \mathbb{R}^{k_j-1}$ con $c_s^{(j)} = \gamma_s^{(j)} - \gamma_{k_j}^{(j)}$, para $1 \leq s \leq k_j - 1$, las funciones aditivas η_j pueden aproximarse por $\sum_{s=1}^{k_j-1} c_s^{(j)} B_s^{(j)}$ y, entonces, el modelo (2.1) puede aproximarse como

$$Y = \mu + \boldsymbol{\beta}^{\mathrm{T}}\mathbf{Z} + \sum_{j=1}^{p} \sum_{s=1}^{k_j-1} c_s^{(j)} B_s^{(j)}(X_j) + u, \tag{2.3}$$

o bien

$$Y = \mu + \boldsymbol{\beta}^{\mathrm{T}}\mathbf{Z} + \mathbf{c}^{\mathrm{T}}\mathbf{V} + u,$$

donde $\mathbf{c} = (\mathbf{c}^{(1)\mathrm{T}} \ldots, \mathbf{c}^{(p)\mathrm{T}})^{\mathrm{T}} \in \mathbb{R}^K$, con $K = \sum_{j=1}^{p} k_j - p$ la dimensión efectiva del espacio utilizado para aproximar las componentes aditivas no paramétricas, $\mathbf{V} = (\mathbf{V}^{(1)}(X_1)^{\mathrm{T}}, \ldots, \mathbf{V}^{(p)}(X_p)^{\mathrm{T}})^{\mathrm{T}}$ y $\mathbf{V}^{(j)}(t) = (B_1^{(j)}(t), \ldots, B_{k_j-1}^{(j)}(t))^{\mathrm{T}}$, para $1 \leq j \leq p$. Este procedimiento lleva a tener una matriz de diseño bien condicionada. Luego, los residuos serán:

$$r_i(a, \mathbf{b}, \mathbf{c}) = Y_i - a - \mathbf{b}^{\mathrm{T}}\mathbf{Z}_i - \sum_{j=1}^{p} \sum_{s=1}^{k_j-1} c_s^{(j)} B_s^{(j)}(X_{ij}) = Y_i - a - \mathbf{b}^{\mathrm{T}}\mathbf{Z}_i - \mathbf{c}^{\mathrm{T}}\mathbf{V}_i\,, \tag{2.4}$$

con $\mathbf{V}_i = (\mathbf{V}^{(1)}(X_{i1})^{\mathrm{T}}, \ldots, \mathbf{V}^{(p)}(X_{ip})^{\mathrm{T}})^{\mathrm{T}}$ y, entonces, los estimadores de mínimos cuadrados se obtendrán, bajo este nuevo modelo (2.3) que aproxima el modelo (2.1), como

$$(\widehat{\mu}, \widehat{\boldsymbol{\beta}}, \widehat{\mathbf{c}}) = \underset{a,\mathbf{b},\mathbf{c}}{\operatorname{argmin}} \sum_{i=1}^{n} r_i^2(a, \mathbf{b}, \mathbf{c}) \tag{2.5}$$

con $\widehat{\mathbf{c}} = (\widehat{\mathbf{c}}^{(1)\mathrm{T}}, \ldots, \widehat{\mathbf{c}}^{(p)\mathrm{T}})^{\mathrm{T}}$ y $\widehat{\mathbf{c}}^{(j)} = (\widehat{c}_1^{(j)}, \ldots, \widehat{c}_{k_j-1}^{(j)})^{\mathrm{T}}$, $1 \leq j \leq p$. Finalmente, para cada $j = 1, \ldots, p$, el estimador de la función aditiva η_j se obtendrá como

$$\widehat{\eta}_j(x) = \sum_{s=1}^{k_j-1} \widehat{c}_s^{(j)} B_s^{(j)}(x)\,.$$

Capítulo 3

Selección de variables

En la práctica, generalmente son recopiladas muchas variables explicativas con el fin de atenuar posibles sesgos a la hora de explicar una variable respuesta. Decidir qué covariables mantener en el modelo y qué variables no son informativas es de suma importancia en la práctica, aunque siempre resulta en una tarea demandante del análisis de datos. El desarrollo de procedimientos de selección de variables ha crecido rápidamente desde la década de 1970. Akaike (1973) propuso un criterio de información Akaike (AIC) con el objetivo de proporcionar un método que permita comparar diferentes modelos y elegir el que mejor explicara los datos, sin sobreajustar el modelo. Es decir, se busca el equilibrio entre la bondad del ajuste del modelo y su complejidad. Hocking (1976) brindó una revisión exhaustiva de los primeros desarrollos del procedimiento de selección de variables. Schwarz (1978) desarrolló el criterio de información Bayesiana (BIC), y Foster y George (1994) el criterio de inflación del riesgo (RIC). Con esos criterios, y con el objetivo de poder seleccionar variables, los estadísticos usualmente usan métodos del tipo "subset selection" y "stepwise selection". Aunque ellos, en la práctica, resultan muy útiles, estos procedimientos de selección ignoran errores estocásticos inherentes a las etapas de la selección de variables y, por lo tanto, las propiedades teóricas de estos procedimientos son difíciles de entender.

En cuanto a la estimación de mínimos cuadrados en modelos esparsos, a menudo no estamos completamente satisfechos con sus resultados y existen dos razones principales para esto. La primera se refiere a la precisión de la predicción: las estimaciones de mínimos cuadrados suelen tener un sesgo bajo pero presentan una gran varianza y, en situaciones donde el modelo es esparso, la precisión de la predicción puede mejorarse si se reducen o fijan en cero algunos coeficientes. Esto implica, en definitiva, sacrificar una pequeña cantidad de sesgo con el fin de reducir la varianza de los valores predichos, lo que puede resultar en una mejora

en la precisión general del modelo. La segunda razón es la interpretación de los resultados: cuando contamos con una gran cantidad de variables, suele ser de interés identificar un subconjunto más pequeño de variables que exhiba los efectos más fuertes. Esto permite obtener una visión más clara y manejable del modelo, aunque en ocasiones implique sacrificar algunos detalles menos significativos. Esta necesidad de simplificar la interpretación y mejorar la precisión predictiva da paso a los procedimientos de selección y regularización que discutiremos más adelante.

3.1. Método: "mejor subconjunto" (*best subset*)

El procedimiento de selección de variables denominado "mejor subconjunto" (*best subset*) es ampliamente utilizado en modelos de regresión, ya que permite identificar el subconjunto óptimo de variables dentro de un conjunto dado. Este enfoque evalúa todos los subconjuntos posibles de variables y selecciona aquel que minimiza un criterio de ajuste, como el error cuadrático medio (MSE), el coeficiente de determinación ajustado (R^2_{ajustado}), el criterio de información de Akaike (AIC) o el criterio de información Bayesiana (BIC). Es considerado uno de los métodos más confiables para la selección de variables, ya que explora exhaustivamente todas las combinaciones posibles de variables, lo que asegura que no se pase por alto ninguna combinación potencialmente significativa.

Según James et al. (2013), este método resulta particularmente útil en problemas donde la identificación precisa de las variables relevantes es fundamental, debido a que examina exhaustivamente todas las combinaciones posibles. Esta característica asegura que ninguna combinación potencialmente significativa quede descartada a priori, lo que lo convierte en una herramienta confiable para la selección de variables en problemas de complejidad moderada.

Supongamos que se tienen q variables explicativas. El algoritmo que permite el cálculo del mejor subconjunto $\mathcal{S}^*$ es el siguiente:

Paso 1. Para cada $k \in \{0, 1, \ldots, q\}$, se consideran todos los subconjuntos $\mathcal{S}$ de covariables de tamaño k.

Paso 2. Para cada subconjunto $\mathcal{S}$, se ajusta el modelo de regresión.

Paso 3. Se calcula el valor del criterio (por ejemplo, MSE, AIC, BIC) seleccionado para cada modelo ajustado.

Paso 4. Se elige el subconjunto $\mathcal{S}^*$ que optimiza el criterio.

A pesar de sus ventajas, la selección *best subset* presenta una desventaja crítica: su alto costo computacional. Dado que para un conjunto de q covariables, el número de subconjuntos posibles es 2^q, el costo de evaluar todos los subconjuntos se vuelve inviable cuando q es grande. Esto hace que el método sea impracticable en problemas con muchas variables. Además, el procedimiento es extremadamente variable debido a su naturaleza discreta, lo que puede generar resultados inestables, particularmente cuando los datos contienen ruido o presentan colinealidad entre las variables (ver Breiman, 1996, Fan y Li, 2001). Frente a la multicolinealidad, la inestabilidad del método podría seleccionar subconjuntos de variables altamente correlacionadas, lo que llevaría a una estimación poco confiable.

Hastie et al. (2009) observaron que, debido a su complejidad computacional, en muchos casos es más eficiente recurrir a otros métodos de selección más simples, como la selección paso a paso (selección *stepwise*) que, aunque menos exhaustivos, permiten una identificación más rápida de variables relevantes.

3.2. Métodos de inclusión y exclusión de variables

Los métodos de inclusión y exclusión de variables son utilizados para identificar subconjuntos de variables relevantes en un modelo de regresión. A diferencia del procedimiento "mejor subconjunto" que evalúa exhaustivamente todas las combinaciones posibles de variables, estos métodos realizan este proceso de forma iterativa, agregando o eliminando en cada paso según un criterio predeterminado. Esto hace que la selección de variables sea más eficiente desde el punto de vista computacional, aunque, como se verá, conlleva ciertas limitaciones.

La característica principal de estos métodos es su flexibilidad: pueden incorporar nuevas variables al modelo (selección *forward*), eliminar variables no significativas (eliminación *backward*) o combinar ambos enfoques en un único proceso (selección *stepwise*), lo que les permite ajustarse de manera dinámica a las características del conjunto de datos. Si bien estos procedimientos son particularmente útiles cuando se busca un modelo relativamente parsimonioso y cuentan con una eficiencia computacional mayor que el método de selección *best subset*, presentan desventajas como la posibilidad de sobreajuste y la dependencia en el criterio de selección utilizado. En definitiva, el modelo final puede estar sesgado ya que la selección de variables depende de las correlaciones entre ellas y del orden en que se agregan o eliminan. A pesar de estas limitaciones, la selección "paso a paso" sigue siendo una herramienta valiosa en la práctica de la estadística y la modelización, especialmente en problemas con grandes cantidades de predictores y cuando no se

dispone de un conocimiento previo sobre la relevancia de cada variable.

3.2.1. Método: "hacia adelante" (selección *forward*)

La selección "hacia adelante" (selección *forward*) es un procedimiento secuencial que comienza con un modelo vacío (sin variables) y agrega iterativamente la variable que mejora un criterio de ajuste, como puede ser el AIC o el R^2 ajustado. Este enfoque es particularmente útil cuando se tiene una gran cantidad de variables y se desea encontrar un modelo parsimonioso sin necesidad de evaluar todas las combinaciones posibles. Según Draper (1998), la selección hacia adelante es más eficiente en términos computacionales que *best subset* y es capaz de identificar rápidamente un subconjunto de variables que mejora más significativamente el ajuste del modelo, lo que lo convierte en una opción bastante popular en la práctica.

El algoritmo para obtener el subconjunto de variables mediante el procedimiento de selección "hacia adelante" es el siguiente:

Paso 1. Iniciar con el modelo nulo, sin covariables.

Paso 2. Para cada variable: incluirla, ajustar el modelo y calcular el criterio seleccionado.

Paso 3. Elegir la variable que optimiza el criterio e incluirla en el modelo.

Paso 4. Repetir el proceso, evaluando las variables restantes, hasta que no se observe una mejora significativa en el criterio.

Una limitación importante de la selección *forward* es que, debido a su naturaleza secuencial, puede ser sensible a las primeras selecciones. Si la primera variable seleccionada no es la más relevante, el modelo puede quedar atrapado en un subóptimo. Además, este método no evalúa interacciones entre variables, lo que puede ser una desventaja si se consideran interacciones importantes en el modelo.

3.2.2. Método: "hacia atrás" (eliminación *backward*)

El procedimiento de eliminación "hacia atrás" (eliminación *backward*) es un método que comienza con un modelo completo que incluye todas las variables disponibles y elimina, iterativamente, la variable menos significativa en función de un criterio preestablecido. Esta estrategia es útil cuando se parte de un modelo con

todas las variables disponibles y se desea simplificar el modelo eliminando aquellas menos importantes.

El algoritmo para obtener un subconjunto de variables mediante el procedimiento de eliminación *backward* es el siguiente:

Paso 1. Iniciar la estimación con el modelo completo.

Paso 2. Para cada variable: excluirla, ajustar el modelo y calcular el criterio seleccionado.

Paso 3. Eliminar la variable cuya exclusión mejora más el criterio.

Paso 4. Repetir el proceso hasta que no se observe una mejora significativa en el criterio.

Una desventaja del procedimiento de eliminación *backward* es que este mecanismo puede ser ineficiente si el número de variables es grande, ya que el modelo comienza con todas las variables, lo que requiere más pasos computacionales. Además, el método *backward* al igual que el *forward*, puede ser problemático en presencia de multicolinealidad, ya que puede eliminar variables importantes que están correlacionadas con otras variables significativas. Además, al igual que sucede con el método *forward*, tiene limitaciones por su inestabilidad y sólo se evalúan mejoras locales. Un punto interesante de observar es que este procedimiento no es aplicable si $q > n$ ya que requiere iniciar con el modelo completo. En este escenario, al comenzar con todas las variables, el modelo inicial no puede ajustarse adecuadamente mediante métodos como los mínimos cuadrados ordinarios, ya que el sistema de ecuaciones subyacente es indeterminado.

3.2.3. Método: "paso a paso" (*stepwise*)

El método de selección *stepwise* combina elementos de la selección *forward* y de la selección *backward*. En cada paso, agrega o elimina variables de manera secuencial dependiendo de si su inclusión o eliminación mejora el criterio de ajuste. Este método es popular por su flexibilidad y su capacidad de manejar tanto la incorporación de nuevas variables como la eliminación de las menos relevantes, lo que lo convierte en una opción versátil para el ajuste de modelos. James et al. (2013) explican que la selección *stepwise* es útil cuando se desea encontrar un modelo equilibrado entre parsimonia y ajuste, especialmente cuando no se conoce de antemano qué variables son las más relevantes.

El algoritmo del método de selección *stepwise* es el siguiente:

Paso 1. Comenzar con el modelo nulo o completo según criterio del investigador.

Paso 2. En cada paso realizar alguna de las siguientes acciones:

- Para cada variable que esté en el modelo: excluirla, ajustar el modelo y calcular el criterio seleccionado.
- Para cada variable que no esté en el modelo: incluirla, ajustar el modelo y calcular el criterio seleccionado.

Paso 3. Elegir la acción (agregar o eliminar) que optimice el criterio seleccionado.

Paso 4. Repetir el proceso hasta que no se observe una mejora significativa en el criterio.

Aunque la selección *stepwise* ofrece una mayor flexibilidad que otros métodos secuenciales, también presenta algunas limitaciones. Una de las críticas más comunes es que este método puede ser inestable, ya que el modelo seleccionado depende en gran medida del orden en que se agregan o eliminan las variables. James et al. (2013) apuntan que este enfoque puede conducir a la selección de un modelo sobreajustado (overfitting), especialmente si no se controla adecuadamente el número de variables incluidas en el modelo. Otra limitación es la posibilidad de que el algoritmo entre en un ciclo, repitiendo secuencias de inclusión y exclusión de variables sin alcanzar un modelo final estable. Esto puede ocurrir, por ejemplo, si el criterio utilizado para evaluar los modelos (como AIC o BIC) presenta pequeñas variaciones entre modelos sucesivos. En estos casos, una variable puede ser agregada en una iteración y eliminada en la siguiente, generando oscilaciones entre modelos similares. Si el algoritmo no incorpora una regla explícita para detectar ciclos o un número máximo de iteraciones, esta situación puede derivar en una no convergencia del procedimiento. Por este motivo, muchas implementaciones prácticas imponen restricciones que previenen el regreso a configuraciones ya evaluadas o detienen el proceso cuando se detecta estabilización en el conjunto de variables seleccionadas.

En conclusión, los métodos clásicos de selección de variables, como el método de selección *best subset* y los procedimientos de inclusión y exclusión de variables (que incluyen los procedimientos de selección *forward*, *backward* y *stepwise*), han sido ampliamente utilizados debido a su simplicidad conceptual, facilidad de implementación y capacidad para identificar subconjuntos relevantes de variables. Estos enfoques permiten explorar el espacio de modelos de manera exhaustiva (como lo hace *best subset*) o secuencial (como lo hacen los métodos de inclusión y

exclusión), logrando buenos resultados en problemas con un número moderado de variables explicativas.

Sin embargo, presentan importantes limitaciones en escenarios más complejos, como aquellos caracterizados por alta dimensionalidad ($q > n$), multicolinealidad severa o la necesidad de balancear parsimonia y capacidad predictiva. Entre las principales desventajas de estos métodos se encuentran:

- tendencia al sobreajuste: al no incorporar mecanismos explícitos de penalización, estos métodos tienden a seleccionar modelos excesivamente complejos, especialmente cuando hay ruido en los datos o muchas variables irrelevantes.
- inestabilidad: la selección discreta de variables hace que pequeñas perturbaciones en los datos puedan alterar significativamente el subconjunto seleccionado, afectando la confiabilidad de las predicciones.
- altos costos computacionales: en el caso de la selección *best subset*, se tiene un gran costo computacional cuando el número de variables es grande. Aunque los métodos secuenciales son más eficientes, sacrifican exhaustividad y pueden pasar por alto combinaciones importantes de variables.

Estas limitaciones resaltan la necesidad de adoptar enfoques avanzados que integren selección de variables con estabilidad, eficiencia computacional y capacidad predictiva. Los métodos de regularización, que desarrollaremos en la siguiente sección, abordan estos desafíos mediante la incorporación de términos de penalización en la función objetivo. Esto no sólo reduce el riesgo de sobreajuste, sino que también fomenta la parsimonia de los modelos.

3.3. Métodos de regularización

En el contexto de la selección de modelos y de estimación en regresión, los métodos de regularización (o *shrinkage methods*) juegan un papel central. Estos métodos buscan alcanzar un compromiso entre la capacidad predictiva del modelo y su complejidad, introduciendo una penalidad que controla la magnitud de los coeficientes y permite seleccionar variables de manera automática, es decir, este tipo de métodos selecciona variables y estima coeficientes simultáneamente. En el contexto de regresión lineal suponiendo el intercept $\mu = 0$ por simplicidad, es decir, bajo el modelo

$$Y = \boldsymbol{\beta}^{\mathrm{T}}\mathbf{Z} + u,$$

la formulación general de un estimador basado en mínimos cuadrados con término de regularización puede expresarse como:

$$\widehat{\boldsymbol{\beta}} = \underset{\mathbf{b}\in\mathbb{R}^q}{\operatorname{argmin}} \sum_{i=1}^{n} \left(Y_i - \sum_{j=1}^{q} b_j Z_{ij} \right)^2 + P_\lambda(\mathbf{b}), \tag{3.1}$$

donde $P_\lambda(\mathbf{b})$ es un término de penalización que depende de λ, un hiperparámetro que controla la intensidad de la regularización. Este término de penalización impone restricciones sobre los coeficientes $\mathbf{b} \in \mathbb{R}^q$, buscando controlar su magnitud, la cantidad de coeficientes diferentes de cero (esparsidad) o ambas características, dependiendo del tipo de penalidad aplicada.

El hiperparámetro λ desempeña un rol crucial en los métodos de regularización. Un λ pequeño resulta en un ajuste más cercano al modelo clásico de mínimos cuadrados, mientras que un λ grande induce mayor penalización, reduciendo la complejidad del modelo y forzando algunos coeficientes hacia cero. La elección de λ se realiza, típicamente, mediante procedimientos de validación cruzada, buscando minimizar el error de predicción en datos no observados (ver Hastie et al., 2015). Otros criterios utilizados con frecuencia como el criterio de Información de Akaike (AIC), propuesto por Akaike (1973), o el criterio de Información Bayesiano (BIC), propuesto por Schwarz (1978), son herramientas comúnmente utilizadas para la selección de λ al balancear la capacidad predictiva del modelo con su complejidad. La elección adecuada de λ es crucial para evitar el sobreajuste, particularmente en contextos de alta dimensionalidad, donde la complejidad del modelo puede aumentar significativamente.

Una observación final, aunque no por ello menos relevante, es que la hipótesis $\mu = 0$ no implica pérdida de generalidad, ya que en la práctica es común centrar la variable respuesta y las covariables antes del ajuste. Además en muchas situaciones se estima preliminarmente el intercept, tal como se plantea en Boente y Martínez (2023). En caso contrario, el modelo puede reformularse considerando $Y_i - \widehat{\mu}$ como variables respuesta, y el procedimiento de estimación continúa de manera análoga.

3.3.1. Estimador Ridge

Un ejemplo particular de un estimador que utiliza un mecanismo de regularización es el estimador Ridge, que fue definido previamente en (2.2). Este estimador corresponde a un caso particular de la formulación general (3.1), donde el término de penalización utilizado es la penalidad L_2, es decir, $P_\lambda(\mathbf{b}) = \lambda \sum_{j=1}^{q} b_j^2$. Como ya hemos visto, el estimador Ridge busca controlar la magnitud de los coeficientes $\mathbf{b}$, reduciendo el impacto de la multicolinealidad y mejorando la estabilidad de las estimaciones, aunque no realiza selección de variables. Tal como se menciona en Fan

y Li (2001), para que una función de penalidad de origen a estimadores esparsos, es decir, con la capacidad de seleccionar variables, es necesario que la función de penalidad sea singular en el origen. Sin embargo, el estimador Ridge sirve como base conceptual para desarrollar otros enfoques que combinan regularización con selección de variables.

3.3.2. Estimador LASSO

Entre los métodos de regularización más conocidos se encuentra el estimador LASSO (*Least Absolute Shrinkage and Selection Operator*), introducido por Tibshirani (1996), y resulta ser uno de los más populares para la selección de variables y estimación simultánea en modelos lineales. Este método incorpora una penalidad L_1 en la función de pérdida de mínimos cuadrados, es decir, en la forma general (3.1), la función de pérdida es $P_\lambda(\mathbf{b}) = \lambda \sum_{j=1}^{q} |b_j|$. En definitiva, el estimador LASSO se define como:

$$\widehat{\boldsymbol{\beta}}^{\text{LASSO}} = \underset{\mathbf{b}\in\mathbb{R}^q}{\operatorname{argmin}} \sum_{i=1}^{n} \left(Y_i - \sum_{j=1}^{q} b_j Z_{ij} \right)^2 + \lambda \sum_{j=1}^{q} |b_j|.$$

Esta penalidad, a diferencia de Ridge, induce esparsidad en los coeficientes, promoviendo que algunos de ellos sean exactamente cero, lo que mejora la interpretabilidad del modelo y permite manejar situaciones de alta dimensionalidad donde el número de variables puede superar ampliamente al de observaciones (ver Tibshirani, 1996, para más información). Además, LASSO controla el sobreajuste al restringir la magnitud de los coeficientes y es computacionalmente eficiente, como señalan Hastie et al. (2015), lo que contribuye a su amplia popularidad. Sin embargo, en presencia de alta correlación entre variables, este método tiende a seleccionar una de ellas arbitrariamente, ignorando al resto. Además, puede no ser consistente en la selección de variables cuando el número de variables es muy grande en comparación con el tamaño de la muestra, tal como se menciona en Zou (2006). Finalmente Fan y Li (2001) señalan que el sesgo creado por este estimador es notablemente grande.

3.3.3. Estimador ADALASSO

Para abordar las limitaciones del estimador LASSO, Zou (2006) propuso utilizar una penalidad aleatoria, dando origen al estimador LASSO adaptativo, denotado por ADALASSO, y que resulta en una extensión del LASSO que mejora la consistencia en la selección de variables al introducir pesos adaptativos en la

penalización. Considerando una penalización de la forma $P_\lambda(\mathbf{b}) = \lambda \sum_{j=1}^{q} w_j |b_j|$ en (3.1), la función objetivo de ADALASSO se expresa como:

$$\widehat{\boldsymbol{\beta}}^{\text{ADALASSO}} = \underset{\mathbf{b}\in\mathbb{R}^q}{\operatorname{argmin}} \sum_{i=1}^{n} \left(Y_i - \sum_{j=1}^{q} b_j Z_{ij} \right)^2 + \lambda \sum_{j=1}^{q} w_j |b_j|,$$

donde los pesos se calculan a partir de un estimador inicial $\widetilde{\boldsymbol{\beta}} = (\widetilde{\beta}_1, \ldots, \widetilde{\beta}_q)^{\mathrm{T}}$ consistente de $\boldsymbol{\beta}$ de la siguiente manera:

$$w_j = \frac{1}{|\widetilde{\beta}_j|^\gamma}, \quad \gamma > 0.$$

y donde entendemos que $|b_j|/|\widetilde{\beta}_j|^\gamma = \infty$ si $|\widetilde{\beta}_j| = 0$ y $|b_j| \neq 0$, mientras que $|b_j|/|\widetilde{\beta}_j|^\gamma = 0$ si $|b_j| = |\widetilde{\beta}_j| = 0$. Esta adaptación permite penalizar menos a los coeficientes grandes, facilitando la identificación de las variables verdaderamente relevantes.

3.3.4. Estimadores Bridge y Elastic Net

Otra función penalidad es la propuesta por Frank y Friedman (1993) que da origen al estimador Bridge: generaliza a Ridge y LASSO mediante una función de penalidad del tipo L_q, con $q > 0$, que permite controlar el grado de regularización. En particular, cuando $q = 1$, Bridge coincide con LASSO, mientras que para $q = 2$, coincide con Ridge. La forma general de la regresión Bridge es:

$$\widehat{\boldsymbol{\beta}}^{\text{BRIDGE}} = \underset{\mathbf{b}\in\mathbb{R}^q}{\operatorname{argmin}} \sum_{i=1}^{n} \left(Y_i - a - \sum_{j=1}^{q} b_j Z_{ij} \right)^2 + \lambda \sum_{j=1}^{q} |b_j|^q.$$

Si bien para $0 < q < 1$ la penalización Bridge induce una mayor esparsidad que LASSO, una desventaja para estos valores de q es que la función objetivo deja de ser convexa, lo cual complica la optimización y puede derivar en soluciones múltiples o inestables. Por otro lado, para $q > 1$ la función de penalización se vuelve más suave y deja de inducir esparsidad, lo que impide la selección automática de variables: los coeficientes se reducen en magnitud, pero no se anulan exactamente.

Por otro lado, Zou y Hastie (2005) introdujeron el estimador Elastic Net, que combina la penalidad L_1 referente a LASSO y la penalidad L_2 referente a Ridge para abordar la selección de variables en presencia de alta correlación. Su formulación es la siguiente:

$$\widehat{\boldsymbol{\beta}}^{\text{ELASTIC NET}} = \underset{\mathbf{b}\in\mathbb{R}^q}{\operatorname{argmin}} \sum_{i=1}^{n} \left(Y_i - \sum_{j=1}^{q} b_j Z_{ij} \right)^2 + \lambda \left[\alpha \sum_{j=1}^{q} |b_j| + \frac{1-\alpha}{2} \sum_{j=1}^{q} b_j^2 \right],$$

donde el parámetro $\alpha \in [0,1]$ permite ponderar la contribución de cada penalización: cuando $\alpha = 0$ coincide con el estimador Ridge, cuando $\alpha = 1$, se obtiene LASSO, y para valores intermedios, una combinación que busca equilibrar la capacidad de selección de variables del estimador LASSO con la estabilidad frente a colinealidad que proporciona Ridge. Zou y Hastie (2005) mencionan que, a diferencia del estimador LASSO que tiende a seleccionar sólo una variable dentro de un grupo de covariables altamente correlacionadas, la penalidad Elastic Net permite la selección conjunta de dichas variables gracias a la combinación de penalizaciones L_1 y L_2. Esta propiedad, conocida como "efecto de agrupamiento", resulta especialmente útil en presencia de colinealidad. Además, hereda la esparsidad de LASSO y la estabilidad de Ridge. Sin embargo, introduce dos hiperparámetros que requieren ser calibrados, lo que puede aumentar la complejidad en la selección de modelos. Adicionalmente, Elastic Net puede generar modelos menos esparsos que LASSO puro, lo que reduce su interpretabilidad en ciertos casos. También es sensible al peso relativo de las penalidades controlado por el parámetro $\alpha \in [0,1]$, lo que agrega una capa adicional de dificultad para garantizar resultados óptimos, tal como discuten Hastie et al. (2015). Si bien la penalización L_2 incorporada en Elastic Net permite estabilizar la selección en presencia de alta correlación, puede aumentar el sesgo de los estimadores, en especial cuando la verdadera solución es altamente esparsa. Esta limitación ha motivado el desarrollo de penalizaciones no convexas, como SCAD y MCP, que buscan reducir el sesgo y mejorar la capacidad de selección de variables.

3.3.5. Estimador con penalidad MCP

La función de penalidad MCP (*Minimax Concave Penalty*), desarrollada por Zhang (2010), es una penalización no convexa que tiene como objetivo reducir el sesgo de estimación en los coeficientes grandes, una de las principales limitaciones de las penalidades convexas como L_q (con $q \geq 1$). La penalidad MCP se define como:

$$\begin{aligned} P_{\text{MCP},\lambda}(\mathbf{b}) &= \sum_{j=1}^{q} \left[\lambda \int_0^{|b_j|} \left(1 - \frac{x}{\gamma\lambda}\right)^+ dx \right] \\ &= \sum_{j=1}^{q} p_\lambda(b_j), \quad \gamma > 1, \end{aligned}$$

con

$$p_\lambda(b_j) = \left[\lambda \int_0^{|b_j|} \left(1 - \frac{x}{\gamma\lambda}\right)^+ dx \right],$$

donde $(x)^+$ representa la función que toma el valor de x si $x > 0$ y 0 en caso contrario. En términos prácticos, $P_{\text{MCP},\lambda}$ penaliza los coeficientes pequeños de manera similar a LASSO, pero reduce significativamente la penalización cuando es aplicada a coeficientes grandes. De forma explícita, cada término $p_\lambda(b_j)$ de la penalidad puede expresarse como:

$$p_\lambda(b_j) = \begin{cases} \lambda|b_j| - \frac{b_j^2}{2\gamma}, & \text{si } |b_j| \leq \gamma\lambda, \\ \frac{\gamma\lambda^2}{2}, & \text{si } |b_j| > \gamma\lambda. \end{cases}$$

Esta estructura de MCP permite una transición suave entre un comportamiento similar al de LASSO para coeficientes pequeños, favoreciendo la esparsidad, y una penalización que se reduce progresivamente a cero para coeficientes suficientemente grandes. De este modo, MCP evita el sobrepenalizamiento de variables relevantes, reduciendo el sesgo en la estimación.

Bajo ciertas condiciones, el estimador que utiliza la penalidad MCP cumple con la propiedad del oráculo, identificando correctamente las variables relevantes y asignando coeficientes exactamente iguales a cero para las irrelevantes, tal como se menciona en Zhang (2010). Sin embargo, la no convexidad de MCP introduce desafíos adicionales, como la posibilidad de múltiples óptimos locales, pudiendo complicar la optimización. Esta penalidad, además, agrega un hiperparámetro adicional: γ. Mientras que el hiperparámetro λ controla la intensidad de la penalización, el hiperparámetro γ regula la transición entre la penalización para coeficientes pequeños y la ausencia de penalización para coeficientes grandes. La elección adecuada de γ es crucial, ya que ajusta la sensibilidad del método y afecta directamente su estabilidad y capacidad de predicción. Zhang (2010) señala que el estimador que utiliza MCP podría ser muy sensible a la elección de estos hiperparámetros, lo que podría afectar la estabilidad de las soluciones, especialmente en situaciones de alta dimensionalidad o cuando las variables están altamente correlacionadas, y que si bien la utilización la penalidad MCP mejora la estimación de coeficientes grandes, su implementación práctica puede ser más compleja debido a la necesidad de algoritmos especializados.

3.3.6. Estimador con penalidad SCAD

Fan y Li (2001) introdujeron la penalización SCAD (*Smoothly Clipped Absolute Deviation*) como un enfoque innovador para la selección de variables en modelos de regresión. Este método busca superar las limitaciones de las penalizaciones convexas, como LASSO y Ridge, proporcionando un balance entre esparsidad, insesgadez y continuidad en los coeficientes estimados. Su principal ventaja radica en

que combina las fortalezas de los métodos convexos, como la simplicidad y la eficiencia computacional, con la capacidad de reducir el sesgo asociado a estimadores de coeficientes grandes.

Fan y Li (2001) establecieron tres propiedades fundamentales que una penalización óptima debe cumplir:

- Insesgadez: los coeficientes grandes deben estimarse con un sesgo mínimo, lo que evita penalizaciones excesivas que distorsionen los parámetros relevantes.
- Esparsidad: los coeficientes pequeños deben ajustarse a cero automáticamente, reduciendo la complejidad del modelo y facilitando la interpretación.
- Continuidad: la transición entre los coeficientes estimados debe ser suave, garantizando estabilidad en la selección de variables y minimizando la sensibilidad a pequeños cambios en los datos.

Los estimadores con funciones de penalización del tipo L_q, como ocurre con LASSO y Ridge, no satisfacen las tres propiedades requeridas anteriormente. Por el contrario, la penalización SCAD está diseñada para cumplir estas propiedades de manera eficiente. Su estructura no convexa en el intervalo $(0, +\infty)$ y su acotación permiten manejar tanto coeficientes grandes como pequeños de manera eficaz. Tal como describen Fan y Li (2001), su singularidad en el origen permite producir soluciones dispersas (muchos coeficientes estimados son cero), su continuidad garantiza estabilidad en la selección del modelo, y su propiedad de ser acotada por una constante produce estimaciones casi insesgadas para coeficientes grandes. Además, SCAD cumple la propiedad del oráculo, lo que significa que los estimadores funcionan bien si se conoce el submodelo correcto. Dicho de otra manera, bajo condiciones adecuadas, es capaz de identificar correctamente las variables relevantes en el modelo y asignarles coeficientes distintos de cero, mientras que asigna coeficientes cero a las variables irrelevantes. Esta propiedad asegura que SCAD no sólo selecciona variables de manera efectiva, sino que también produce estimaciones casi insesgadas para los coeficientes de las variables seleccionadas, incluso en situaciones de alta dimensionalidad.

La función de penalización SCAD está definida como:

$$P_{\text{SCAD},\lambda}(t) = \begin{cases} \lambda|t| & \text{si } |t| \leq \lambda \\ -\frac{t^2 - 2a\lambda|t| + \lambda^2}{2(a-1)} & \text{si } \lambda < |t| \leq a\lambda \\ \frac{\lambda^2(a+1)}{2} & \text{si } |t| > a\lambda \end{cases},$$

donde $a > 2$ es un hiperparámetro que controla la transición entre las regiones de penalización lineal, cuadrática y constante.

Además de su formulación teórica sólida, SCAD ha demostrado ser altamente versátil en aplicaciones prácticas. Este enfoque ha ganado popularidad y ha sido ampliamente estudiado en distintos contextos. Entre otros, se puede mencionar a Fan y Li (2002) que lo aplicaron para modelos de Cox; Li y Liang (2008) a modelos semiparamétricos; y Liang y Li (2009) a modelos parcialmente lineales con errores de medición. Además, Xie y Huang (2009) estudiaron la selección de variables para modelos parcialmente lineales con un número de divergente de covariables lineales y establecieron consistencia en la selección y normalidad asintótica. Más recientemente, Boente y Martínez (2024) propusieron un procedimiento robusto para modelos parcialmente lineales aditivos, en el que se seleccionan las variables relevantes de la parte lineal del modelo mediante una penalización tipo SCAD.

La Figura 3.1 muestra un gráfico comparativo de las funciones de penalización Ridge (en línea roja), LASSO (en azul), Elastic Net (en violeta), MCP (en naranja) y SCAD (en verde). Para Elastic Net, se ha considerado $\alpha = 0.5$, representando un caso de equilibrio entre Ridge y LASSO. Siguiendo la recomendación de Fan y Li (2001), se ha tomado $a = 3.7$ para la penalización SCAD, mientras que para MCP se ha utilizado $\gamma = 3$, conforme a la sugerencia de Zhang (2010).

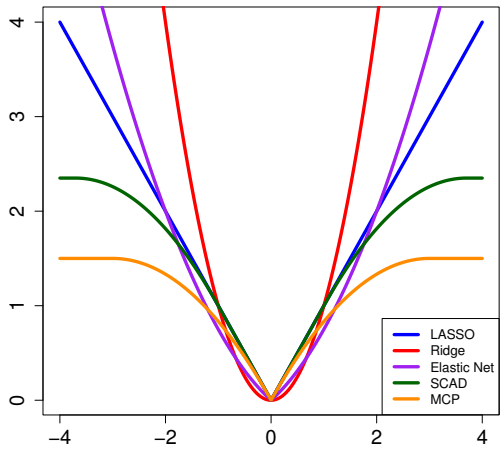

Figura 3.1: Gráfico comparativo de las funciones de penalización Ridge, LASSO, Elastic Net, SCAD y MCP.

Tal como se menciona en Chebi (2019), MCP y SCAD tienen comportamientos similares en el sentido de que ambos intentan corregir la penalización excesiva inherente a LASSO y otros métodos convexos. A diferencia de ellos, ambas pe-

nalizaciones llevan a estimadores continuos y evitan los saltos abruptos en los valores de los coeficientes, lo que resulta en modelos más interpretables y con menor sesgo. Además, bajo ciertas condiciones, tanto MCP como SCAD poseen la propiedad del oráculo, tal como se menciona en Fan y Li (2001) y Zhang (2010), lo que les permite seleccionar el modelo correcto y a la vez obtener estimadores consistentes de los coeficientes verdaderos. A pesar de sus similitudes, MCP y SCAD presentan diferencias notables en sus formulaciones matemáticas. Mientras que la penalización de SCAD crece linealmente para valores pequeños de los coeficientes y luego decrece suavemente hasta estabilizarse, MCP utiliza una penalización que aumenta de forma lineal con pendiente decreciente, volviéndose constante cuando el coeficiente alcanza un umbral específico. Estas diferencias pueden influir en el desempeño de cada método en escenarios prácticos, aunque en esencia ambos comparten el objetivo de reducir el sesgo asociado a las penalizaciones convexas. De hecho, estudios comparativos como los realizados por Zhang (2010) y Fan y Lv (2011) han destacado estas características, subrayando que la elección entre MCP y SCAD puede depender tanto de las propiedades del conjunto de datos como de las preferencias del investigador en términos de complejidad y rendimiento computacional. En nuestro caso, consideramos que las propiedades de SCAD, junto con su capacidad para reducir el sesgo de forma efectiva sin requerir un ajuste computacional complejo, la convierten en una opción adecuada para los objetivos de esta investigación.

3.4. Selección de variables en modelos semiparamétricos esparsos

La idea de aplicar mínimos cuadrados penalizados para la selección de variables en modelos de regresión lineal se ha extendido de forma natural a contextos más generales, como los modelos parcialmente lineales (PLM) y los modelos parcialmente lineales aditivos (PLAM).

Entre los trabajos que seleccionan variables en modelos PLM, podemos mencionar el trabajo de Fan y Li (2004), quienes propusieron un estimador basado en núcleos para la estimación de funciones no paramétricas, combinando este enfoque con la penalización SCAD para la selección de variables. Por su parte, Bunea (2004) introdujo una clase de estimadores tipo sieve basados en mínimos cuadrados penalizados para modelos semiparamétricos, mientras que Bunea y Wegkamp (2004) sugirieron un procedimiento de estimación en dos etapas con un enfoque similar. En el contexto de alta dimensión, Xie y Huang (2009) estudiaron la regresión penalizada con SCAD, utilizando splines polinomiales para estimar la componente

no paramétrica. Ni et al. (2009) propusieron un procedimiento de mínimos cuadrados con doble penalización: una para la obtención de splines suavizados y otra, basada en SCAD, para la selección de variables en la componente lineal. Liu et al. (2018) desarrollaron un método de selección de variables en contextos de alta dimensión extrema, utilizando correlaciones parciales entre la respuesta y las covariables. Por otro lado, Zhao y Xue (2009) abordaron la selección de variables en modelos semiparamétricos parcialmente lineales con coeficientes que varían con ciertas covariables. Para ello, propusieron un método que combina la aproximación mediante $B-$splines para la parte no paramétrica con una penalización LASSO en grupos, aplicada simultáneamente sobre los coeficientes constantes y los coeficientes variables.

Siguiendo esta misma línea, varios trabajos se han enfocado en extender técnicas de selección de variables en la componente lineal a modelos PLAM, sin abordar directamente la parte no paramétrica. Entre ellos, Liu et al. (2011) propusieron un procedimiento de estimación y selección de variables que utiliza aproximaciones por splines para las funciones aditivas, junto con una penalización SCAD aplicada a los coeficientes lineales. Su metodología permite estimar de forma conjunta la componente paramétrica y la no paramétrica, y seleccionar automáticamente las variables relevantes en la parte lineal del modelo. Wei (2012) propusieron aplicar el *adaptive group LASSO*, introducido por Wang y Leng (2008), para seleccionar grupos de variables relevantes de la componente lineal, bajo el supuesto de existencia de agrupamientos en dicha componente (modelo que denominan GPLAM). En su enfoque, la componente no paramétrica se estima mediante splines. Por su parte, Boente y Martínez (2024) desarrollaron un método robusto para la estimación y selección simultánea de variables en modelos parcialmente lineales aditivos, enfocado exclusivamente en la selección de variables en la componente lineal. Su propuesta se basa en una estrategia en dos etapas: primero, se estiman de forma robusta las funciones aditivas, el intercept μ y el parámetro de escala; en la segunda etapa, se aplican MM-estimadores de regresión con penalización sobre los coeficientes lineales, lo cual permite realizar selección de variables en esa componente. En los estudios de simulación y en el análisis de datos reales, se utilizó la penalización SCAD.

Si bien numerosos trabajos han abordado la estimación y selección de variables en modelos lineales, parcialmente lineales y PLAM, bajo el supuesto de esparsidad en el vector de parámetros de regresión, hasta donde sabemos, son escasos los estudios que abordan la selección simultánea de variables en ambas componentes de un modelo PLAM. Entre estos, Du et al. (2012) propusieron un procedimiento que realiza selección simultánea en las componentes lineal y no paramétrica. Su enfoque combina una penalización basada en normas empíricas adaptativas tipo L_2 aplicadas sobre las funciones aditivas, construidas a partir de estimadores ini-

ciales obtenidos mediante splines y penalización Ridge, con un segundo paso de estimación y selección de la componente lineal mediante penalización SCAD. En la misma dirección, Kazemi et al. (2018) propusieron un método en dos etapas para estimación y selección de variables en contextos de alta dimensión, utilizando penalización SCAD y un procedimiento de optimización que minimiza la suma de residuos al cuadrado. Además, desarrollaron un criterio tipo BIC para seleccionar los parámetros de penalización, junto con un mecanismo de validación cruzada con dos particiones para determinar los parámetros de ajuste de la aproximación por $B-$splines. Hu et al. (2015) adoptaron un enfoque bayesiano jerárquico para estimación y selección de modelos en regresión de cuantiles, mientras que Banerjee y Ghosal (2014) propusieron un método bayesiano para modelos aditivos parcialmente lineales generalizados que incorpora LASSO en grupos (ver Yuan y Lin, 2006, para más información sobre LASSO en grupos). Cuando la variable respuesta es una serie temporal, Feng et al. (2018) desarrollaron un procedimiento de estimación y selección de variables basado en el uso de LASSO adaptativo en grupos. Cabe señalar que en todos estos trabajos, las funciones aditivas fueron aproximadas mediante splines, excepto en Feng et al. (2018), donde se utilizaron series ortogonales.

Sin embargo, en ninguno de los trabajos mencionados se propone un procedimiento que combine aproximación por $B-$splines con una estrategia de regularización adaptativa, incorporando además dos criterios basados en BIC para la selección automática de los parámetros de penalización y suavizado, tal como se desarrolla en el presente trabajo de investigación.

Capítulo 4

La propuesta

En este Capítulo, se presenta una nueva propuesta para la estimación y selección simultánea de variables en modelos parcialmente lineales aditivos (PLAM), combinando las ventajas de la flexibilidad no paramétrica y el aporte de los métodos de regularización. Este enfoque aborda la estimación conjunta de las componentes lineales y no paramétricas del modelo, utilizando una base de $B-$splines para aproximar las funciones no paramétricas aditivas, como se detalló en el Capítulo 2.

4.1. Estimación y selección simultánea de variables en modelos PLAM

En términos generales, la propuesta considera la aproximación de cada componente aditiva mediante una base de $B-$splines, lo que permite expresar el modelo PLAM en términos de una regresión lineal extendida. A partir de esta formulación, definimos los estimadores mediante un procedimiento de regularización adaptativo, utilizando la penalidad SCAD para penalizar tanto los coeficientes de la parte lineal como los asociados a las aproximaciones no paramétricas por $B-$splines.

Más precisamente, recordemos que el modelo PLAM viene dado por la expresión (2.1). Tal como se mencionó en la Sección 2.3, al utilizar una base de $B-$splines centrados $\{B_s^{(j)} : 1 \leq s \leq k_j\}$ de orden ℓ_j para la aproximación de la función aditiva η_j, $1 \leq j \leq q$, el modelo (2.1) puede aproximarse por:

$$Y = \mu + \boldsymbol{\beta}^{\mathrm{T}}\mathbf{Z} + \sum_{j=1}^{p}\sum_{s=1}^{k_j-1} c_s^{(j)} B_s^{(j)}(X_j) + u\,.$$

En lugar de estimar mediante mínimos cuadrados, es decir, minimizando el problema (2.5) que sabemos no conduce a un procedimiento de selección de variables, agregaremos un término de penalización que contempla una penalización para la componente lineal y una penalización para la componente aditiva del modelo aproximado. Más precisamente, la propuesta de estimación consiste en hallar $(\widehat{\mu}, \widehat{\boldsymbol{\beta}}, \widehat{\mathbf{c}})$ resolviendo el siguiente problema de minimización

$$(\widehat{\mu}, \widehat{\boldsymbol{\beta}}, \widehat{\mathbf{c}}) = \underset{a\in\mathbb{R}, \mathbf{b}\in\mathbb{R}^q, \mathbf{c}\in\mathbb{R}^K}{\operatorname{argmin}} PL_{n,\boldsymbol{\lambda},\mathbf{k}}(a, \mathbf{b}, \mathbf{c})\,, \tag{4.1}$$

donde

$$PL_{n,\boldsymbol{\lambda},\mathbf{k}}(a, \mathbf{b}, \mathbf{c}) = \frac{1}{n}\sum_{i=1}^{n} r_i^2(a, \mathbf{b}, \mathbf{c}) + p_{\boldsymbol{\lambda},\mathbf{k}}(\mathbf{b}, \mathbf{c}).$$

El término de penalización $p_{\boldsymbol{\lambda},\mathbf{k}}(\mathbf{b}, \mathbf{c})$ está dado por

$$p_{\boldsymbol{\lambda},\mathbf{k}}(\mathbf{b}, \mathbf{c}) = \sum_{s=1}^{q} p_{\lambda_{1,s}}(|b_s|) + \sum_{j=1}^{p} p_{\lambda_{2,j}}(\|\mathbf{c}^{(j)}\|_{H_j})\,, \tag{4.2}$$

donde

- $p_{\lambda_{1,s}}$ $(1 \le s \le q)$ y $p_{\lambda_{2,j}}$ $(1 \le j \le p)$ son funciones de penalidad que deben definirse, y para las cuales sugerimos usar SCAD, basándonos en la discusión previa,
- Los parámetros de penalización $\lambda_{1,s}$ $(1 \le s \le q)$ y $\lambda_{2,j}$ $(1 \le j \le p)$ controlan la esparsidad de las componentes paramétricas y no paramétricas, respectivamente. Por tanto, $\boldsymbol{\lambda} = (\lambda_{1,1}, \ldots, \lambda_{1,q}, \lambda_{2,1}, \ldots, \lambda_{2,p})^{\mathrm{T}}$ representa el vector de penalización,
- $K = \sum_{j=1}^{p} k_j - p$ y $\mathbf{c}^{(j)} = (c_1^{(j)}, \ldots, c_{k_j-1}^{(j)})^{\mathrm{T}}$,
- $\|\mathbf{c}^{(j)}\|_{H_j} = (\mathbf{c}^{(j)\mathrm{T}} H_j \mathbf{c}^{(j)})^{1/2}$, donde H_j es una matriz de dimensión $(k_j - 1) \times (k_j - 1)$ con elementos (s, s') dados por $\int_0^1 B_s^{(j)}(t) B_{s'}^{(j)}(t)\, dt$, y
- los residuos $r_i(a, \mathbf{b}, \mathbf{c})$ están definidos como en (2.4).

Nuevamente, tal como ocurría en la Sección 2.3, una vez obtenido $(\widehat{\mu}, \widehat{\boldsymbol{\beta}}, \widehat{\mathbf{c}})$ mediante este nuevo problema de optimización, los estimadores de las funciones aditivas se definen, para cada $1 \le j \le q$, como

$$\widehat{\eta}_j(x) = \sum_{s=1}^{k_j-1} \widehat{c}_s^{(j)} B_s^{(j)}(x)\,.$$

4.2. Selección del parámetro de penalidad

Como se ha discutido, por ejemplo, en Efron et al. (2004) y Meinshausen y Bühlmann (2010), la selección del parámetro de penalización es crucial cuando se busca ajustar modelos esparsos, ya que este parámetro regula la complejidad del modelo.

Consideremos $\mathbf{k} = (k_1, \ldots, k_p)^{\mathrm{T}}$, el vector que contiene la información del número de términos para aproximar las funciones aditivas, fijo. Nuestro objetivo es seleccionar los parámetros de regularización $\lambda_{1,s}$ (para $1 \leq s \leq q$) y $\lambda_{2,j}$ (para $1 \leq j \leq p$) que determinan la esparsidad de las componentes lineales y no lineales, respectivamente. Desafortunamente, elegir $\boldsymbol{\lambda} = (\lambda_{1,1}, \ldots, \lambda_{1,q}, \lambda_{2,1}, \ldots, \lambda_{2,p})^{\mathrm{T}}$ sobre una grilla de dimensión $q + p$ optimizando algún criterio, como ser AIC o BIC, puede ser computacionalmente muy costoso si $q + p$ es grande. Para resolver este problema, adoptaremos un enfoque adaptativo basado en la idea de Zhang *et al.* (2014) que simplifica el problema a la selección de sólo dos parámetros, λ_1 y λ_2, es decir, reduciendo la búsqueda de los valores óptimos sobre una grilla de dimensión 2. En este enfoque, los parámetros de la penalización son definidos en función de estos parámetros auxiliares como:

$$\lambda_{1,s} = \frac{\lambda_1}{|\widetilde{\beta}_s|} \qquad \text{y } \lambda_{2,j} = \frac{\lambda_2}{\|\widetilde{\mathbf{c}}^{(j)}\|_{H_j}}, \tag{4.3}$$

donde $\widetilde{\beta}_s$ y $\widetilde{\mathbf{c}}^{(j)}$ son estimadores iniciales no penalizados de β_s y $\mathbf{c}^{(j)}$, respectivamente. Luego, para seleccionar el par óptimo $\widetilde{\boldsymbol{\lambda}} = (\lambda_1, \lambda_2)$, minimizaremos el siguiente criterio de tipo BIC sobre una grilla $\mathcal{G} \subset \mathbb{R}^2$:

$$\mathrm{BIC}_{\boldsymbol{\lambda}}(\widetilde{\boldsymbol{\lambda}}) = \log\left(\sum_{i=1}^{n} r_i^2(\widehat{\mu}_{\boldsymbol{\lambda}}, \widehat{\beta}_{\boldsymbol{\lambda}}, \widehat{\mathbf{c}}_{\boldsymbol{\lambda}})\right) + df_c \frac{\log(n)}{n} + df_n \frac{\log(n/K)}{n/K}, \tag{4.4}$$

donde

- $(\widehat{\mu}_{\boldsymbol{\lambda}}, \widehat{\beta}_{\boldsymbol{\lambda}}, \widehat{\mathbf{c}}_{\boldsymbol{\lambda}})$ es la solución al problema de optimización definido en (4.1) para el vector dado $\boldsymbol{\lambda} = (\lambda_{1,1}, \ldots, \lambda_{1,q}, \lambda_{2,1}, \ldots, \lambda_{2,p})^{\mathrm{T}}$ con la relación con $\widetilde{\boldsymbol{\lambda}}$ dada por (4.3),
- $r_i(\widehat{\mu}_{\boldsymbol{\lambda}}, \widehat{\beta}_{\boldsymbol{\lambda}}, \widehat{\mathbf{c}}_{\boldsymbol{\lambda}})$ son los residuos definidos en (2.4) evaluados en $(\widehat{\mu}_{\boldsymbol{\lambda}}, \widehat{\beta}_{\boldsymbol{\lambda}}, \widehat{\mathbf{c}}_{\boldsymbol{\lambda}})$,
- df_c es el número de componentes lineales significativas, y
- df_n el número de componentes aditivas significativas.

En lo que respecta a la elección de los estimadores iniciales, una posibilidad es considerar la propuesta realizada por Boente y Martínez (2023) utilizando la función de pérdida cuadrática, es decir, tomando $\rho(t) = t^2$, siempre que sea computable, o considerar la versión con penalidad del tipo Ridge propuesta en la Sección 2.3 de ese mismo trabajo.

4.3. Selección del parámetro de suavizado

Cualquier enfoque basado en $B-$splines requiere la selección de un conjunto de nodos para definir las funciones de la base. En este sentido, dado que cada función aditiva η_j, $1 \leq j \leq p$, se aproxima mediante una curva $B-$spline, son necesarias p secuencias de nodos, una para cada función aditiva. Los nodos uniformes se utilizan cuando la función, en este caso η_j, no presenta cambios dramáticos en sus derivadas, mientras que los nodos no uniformes son preferidos cuando la función muestra comportamientos muy diferentes en distintas regiones. En este último caso, se emplean nodos basados en cuantiles, es decir, los nodos se obtienen como los cuantiles de la variable explicativa observada, con percentiles uniformemente distribuidos. Sin embargo, Stone (1985) menciona que el número de nodos, k_j asociado a la aproximación por $B-$splines de la función aditiva η_j, es más importante que su ubicación. Por esta razón, nos enfocaremos en seleccionar el número de nodos $k_1, \ldots, k_p$ minimizando un criterio BIC.

En este trabajo, hallaremos el valor del vector $\mathbf{k} = (k_1, \ldots, k_p)^{\mathrm{T}}$, que contiene el número de elementos de la base de $B-$splines para aproximar cada función aditiva, minimizando el criterio BIC propuesto en Boente y Martínez (2023). Consideraremos el caso particular de $\rho(t) = t^2$ para la función que controla los residuos. Luego, para un valor del vector de penalización $\boldsymbol{\lambda}$ fijo, el criterio de BIC propuesto en dicho trabajo se reduce a

$$\mathrm{BIC}_{\mathbf{k}}(\mathbf{k}) = \log\left(\sum_{i=1}^{n} r_i^2(\widehat{\mu}_{\boldsymbol{\lambda},\mathbf{k}}, \widehat{\beta}_{\boldsymbol{\lambda},\mathbf{k}}, \widehat{\mathbf{c}}_{\boldsymbol{\lambda},\mathbf{k}})\right) + \frac{\log(n)}{2n}\sum_{j=1}^{p} k_j, \tag{4.5}$$

donde $r_i(\widehat{\mu}_{\boldsymbol{\lambda},\mathbf{k}}, \widehat{\beta}_{\boldsymbol{\lambda},\mathbf{k}}, \widehat{\mathbf{c}}_{\boldsymbol{\lambda},\mathbf{k}})$, $1 \leq i \leq n$ son los residuos obtenidos al utilizar una base de $B-$splines de dimensión k_j para aproximar η_j (para $1 \leq j \leq p$). Vale la pena observar que cuando $p = 1$, el criterio se reduce al propuesto por He et al. (2002). Además, este criterio tiene en cuenta únicamente la cantidad de nodos dados en $\mathbf{k}$ y puede utilizarse en cualquiera de los casos de ubicación de nodos mencionados anteriormente: cuando los nodos son equiespaciados o bien cuando los nodos están basados en los cuantiles de la variable explicativa.

Un caso particular consiste en utilizar el mismo número de elementos de la base

en cada componente aditiva, es decir, $k = k_j$ para $j = 1, \ldots, p$. Luego, el criterio BIC puede redefinirse como

$$\mathrm{BIC}_{\mathbf{k}}(k) = \log\left(\sum_{i=1}^{n} r_i^2(\widehat{\mu}_{\boldsymbol{\lambda},\mathbf{k}}, \widehat{\beta}_{\boldsymbol{\lambda},\mathbf{k}}, \widehat{\mathbf{c}}_{\boldsymbol{\lambda},\mathbf{k}})\right) + p\frac{\log(n)}{2n}k. \tag{4.6}$$

Tal como se menciona en He et al. (2002) y Boente y Martínez (2023), cuando se consideran splines cúbicos, para minimizar el criterio de BIC definido en (4.5), la $j-$ésima componente de $\mathbf{k}$ puede buscarse en el intervalo $\left[\text{máx}\{n^{1/5}/2, 4\}, 8 + 2n^{1/5}\right]$. Los cálculos numéricos se reducen aún más cuando el tamaño de la base es el mismo para todos los j. En este caso, basta con buscar el primer mínimo local de $\mathrm{BIC}_{\mathbf{k}}(k)$, definido en (4.6), para $\text{máx}\{n^{1/5}/2, 4\} \leq k \leq 8 + 2n^{1/5}$.

4.4. Algoritmo general

Esta sección presenta el algoritmo propuesto para la obtención de estimadores penalizados cuando se considera un procedimiento adaptativo, es decir, cuando las componentes del vector de penalización $\boldsymbol{\lambda}$ se definen utilizando los parámetros auxiliares λ_1 y λ_2, según la expresión (4.3). No obstante, el algoritmo es lo suficientemente flexible como para adaptarse a situaciones en las que el vector de penalización no adopta una forma específica. El Algoritmo 1 describe un procedimiento general para calcular los estimadores, basado en una grilla de valores para el vector de penalización auxiliar $\widetilde{\boldsymbol{\lambda}}$ y el vector de regularización $\mathbf{k}$, junto con los criterios BIC definidos en las expresiones (4.4) y (4.5). Por simplicidad y como ya hemos mencionado, sin pérdida de generalidad, supondremos $\mu = 0$. De esta forma, el problema (4.1), se reduce a hallar $(\boldsymbol{\beta}, \widehat{\mathbf{c}})$ resolviendo el problema de optimización

$$(\widehat{\boldsymbol{\beta}}, \widehat{\mathbf{c}}) = \underset{\mathbf{b}\in\mathbb{R}^q, \mathbf{c}\in\mathbb{R}^K}{\operatorname{argmin}} PL_{n,\boldsymbol{\lambda},\mathbf{k}}(\mathbf{c}, \mathbf{c})\,, \tag{4.7}$$

con

$$PL_{n\boldsymbol{\lambda},\mathbf{k}}(\mathbf{b}, \mathbf{c}) = \frac{1}{n}\sum_{i=1}^{n} r_i^2(\mathbf{b}, \mathbf{c}) + p_{\boldsymbol{\lambda},\mathbf{k}}(\mathbf{b}, \mathbf{c})\,,$$

donde los residuos son $r_i(\mathbf{b}, \mathbf{c}) = Y_i - \mathbf{b}^{\mathrm{T}}\mathbf{Z}_i - \mathbf{c}^{\mathrm{T}}\mathbf{V}_i$ y la función de penalidad $p_{\boldsymbol{\lambda},\mathbf{k}}$ se define como antes.

Algoritmo 1 Algoritmo general

1: Obtener estimadores preliminares $\widehat{\mu}$, $\widehat{\boldsymbol{\beta}}$, $\widehat{\eta}_j$, para $1 \leq j \leq p$. Una posible elección es calcular los estimadores obtenidos con el procedimiento no penalizado propuesto en Boente y Martínez (2023) utilizando como función de pérdida la función cuadrática, tal como desarrollamos en la Sección 2.3, o la versión Ridge propuesta en dicho trabajo.
2: Para $j = 1, \ldots, p$, fijar el orden del spline ℓ_j.
3: Considerar una grilla para $\mathbf{k} = (k_1, \ldots, k_p)^{\mathrm{T}}$ de M elementos: $\mathbf{k}_1, \ldots, \mathbf{k}_M$, y considerar una grilla para $\widetilde{\boldsymbol{\lambda}} = (\lambda_1, \lambda_2)^{\mathrm{T}}$ de N elementos: $\widetilde{\boldsymbol{\lambda}}_1, \ldots, \widetilde{\boldsymbol{\lambda}}_N$.
4: **for** $\ell = 1$ **a** M **do**
5: **for** $j = 1$ **a** N **do**
6: Resolver el problema de optimización

$$(\widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}_j,\mathbf{k}_\ell}, \widehat{\mathbf{c}}_{\boldsymbol{\lambda}_j,\mathbf{k}_\ell}) = \underset{\mathbf{b}\in\mathbb{R}^q, \mathbf{c}\in\mathbb{R}^K}{\operatorname{argmin}} PL_{n,\boldsymbol{\lambda}_j,\mathbf{k}_\ell}(\mathbf{b}, \mathbf{c}), \tag{4.8}$$

donde $(\widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}_j,\mathbf{k}_\ell}, \widehat{\mathbf{c}}_{\boldsymbol{\lambda}_j,\mathbf{k}_\ell})$ es el argumento que resuelve el problema de optimización para $\mathbf{k} = \mathbf{k}_\ell$ y el vector de penalidad $\boldsymbol{\lambda} = \boldsymbol{\lambda}_j$, que se relaciona con el vector de parámetros auxiliares $\widetilde{\boldsymbol{\lambda}}_j$ mediante la expresión (4.3), fijos.
7: Calcular y guardar $\mathrm{BIC}_{\boldsymbol{\lambda}}(\widetilde{\boldsymbol{\lambda}}_j)$.
8: **end for**
9: Resolver

$$\widehat{j} = \underset{j}{\operatorname{argmin}}\, \mathrm{BIC}_{\boldsymbol{\lambda}}(\widetilde{\boldsymbol{\lambda}}_j).$$

Tomamos $\widetilde{\boldsymbol{\lambda}}_{\mathbf{k}_\ell} = \widetilde{\boldsymbol{\lambda}}_{\widehat{j}}$. Luego, $\boldsymbol{\lambda}_{\mathbf{k}_\ell}$ es el vector de parámetros con vector de parámetros auxiliares $\widetilde{\boldsymbol{\lambda}}_{\mathbf{k}_\ell} = (\lambda_{1,\mathbf{k}_\ell}, \lambda_{2,\mathbf{k}_\ell})$, obtenido para $\mathbf{k} = \mathbf{k}_\ell$ fijo.
10: Calcular y guardar $\mathrm{BIC}_{\mathbf{k}}(\mathbf{k}_\ell)$.
11: **end for**
12: Obtener $\widehat{\mathbf{k}}$ como

$$\widehat{\ell} = \underset{\ell}{\operatorname{argmin}}\, \mathrm{BIC}_{\mathbf{k}}(\mathbf{k}_\ell).$$

Luego, $\widehat{\mathbf{k}} = \mathbf{k}_{\widehat{\ell}}$.
13: Los estimadores penalizados obtenidos son $\left(\widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}_{\widehat{\mathbf{k}}},\widehat{\mathbf{k}}}, \widehat{\mathbf{c}}_{\boldsymbol{\lambda}_{\widehat{\mathbf{k}}},\widehat{\mathbf{k}}}\right)$.

Como se puede apreciar, el Algoritmo 1 no incluye cómo resolver el problema de optimización (4.8), el cual aún debe ser descrito. En lo que sigue, se desarrolla un método para resolver la tarea de minimización, considerando $\mathbf{k}$ y $\boldsymbol{\lambda}$ fijos.

Resolución del problema de optimización para $\mathbf{k}$ y $\boldsymbol{\lambda}$ fijos

En lo que sigue, consideraremos funciones de penalidad univariadas p_λ que sean dos veces diferenciables en el intervalo $(0; M)$, con $M > 0$ y satisfaciendo $p_\lambda(0) = 0$. Estas condiciones permiten la aproximación local mediante funciones cuadráticas cercanas al origen e incluyen algunas funciones de penalización comunes, como las penalizaciones L_1, SCAD y MCP.

Para $\boldsymbol{\lambda}_j$ y $\mathbf{k}_\ell$ fijos, queremos resolver el problema de optimización (4.8) cuya función objetivo, para esos valores de los parámetros, queda definida como

$$PL_{n,\boldsymbol{\lambda}_j,\mathbf{k}_\ell}(\mathbf{b},\mathbf{c}) = \frac{1}{n}\sum_{i=1}^{n} r_i^2(\mathbf{b},\mathbf{c}) + p_{\boldsymbol{\lambda}_j,\mathbf{k}_\ell}(\mathbf{b},\mathbf{c})\,, \tag{4.9}$$

con $p_{\boldsymbol{\lambda}_j,\mathbf{k}_\ell}$ definida en (4.2), es decir, como suma de funciones de penalidad univariadas tanto para la componente lineal como para la componente aditiva del modelo.

Por simplicidad, supongamos que la función de penalización univariada utilizada p_λ es la misma tanto para la parte lineal como para la parte aditiva del modelo. Para evitar una notación cargada, escribimos $p = p_{\lambda_{1,s}}$ y $p = p_{\lambda_{2,j}}$ según corresponda. Las hipótesis establecidas para p garantizan que, si t_0 es un punto inicial cercano a 0, entonces existe una función cuadrática $q(t)$ simétrica alrededor de 0 tal que $q(t_0) = p(|t_0|)$ y $q'(t_0) = p'(|t_0|)\,\mathrm{sgn}(t_0)$. Usando la simetría, tenemos que $q(t) = a + bt^2$. Luego, las condiciones sobre $q(t_0)$ y $q'(t_0)$ implican que

$$\begin{cases} a + b|t_0|^2 = p(|t_0|), \\ 2bt_0 = p'(|t_0|)\mathrm{sgn}(t_0) \end{cases} \quad \Leftrightarrow \quad \begin{cases} a + b|t_0|^2 = p(|t_0|) \\ 2b|t_0| = p'(|t_0|) \end{cases} .$$

De la segunda ecuación, obtenemos $b = p'(|t_0|)/2|t_0|$ y, reemplazando en la primera ecuación, obtenemos $a = p(|t_0|) - p'(|t_0|)/2|t_0|t_0^2$. Finalmente, la aproximación cuadrática es

$$q(t) = p(|t_0|) - \frac{p'(|t_0|)}{2|t_0|}t_0^2 + \frac{p'(|t_0|)}{2|t_0|}t^2 = p(|t_0|) + \frac{p'(|t_0|)}{2|t_0|}(t^2 - t_0^2).$$

De este modo, dado $(\mathbf{b}_0, \mathbf{c}_0)$ cercano al minimizador de (4.8) y $(\mathbf{b}, \mathbf{c})$ cercano a $(\mathbf{b}_0, \mathbf{c}_0)$ con $|b_{0s}| > 0$ y $\|\mathbf{c}_0^{(j)}\|_{H_j} > 0$ para $s = 1, \ldots, q$ y $j = 1, \ldots, p$, se tiene que

$$p(|b_s|) \approx p(|b_{0s}|) + \frac{p'(|b_{0s}|)}{2|b_{0s}|}(b_s^2 - b_{0s}^2)$$

y

$$p(\|\mathbf{c}^{(j)}\|_{H_j}) \approx p(\|\mathbf{c}_0^{(j)}\|_{H_j}) + \frac{p'(\|\mathbf{c}_0^{(j)}\|_{H_j})}{2\|\mathbf{c}^{(j)}\|_{H_j}}(\|\mathbf{c}^{(j)}\|_{H_j}^2 - \|\mathbf{c}_0^{(j)}\|_{H_j}^2).$$

Recordando que $\|\mathbf{c}^{(j)}\|^2_{H_j} = (\mathbf{c}^{(j)})^{\mathrm{T}} H_j \mathbf{c}^{(j)}$ y llamando $K^* = q + K - p$, definimos

$$\boldsymbol{\Sigma}(\mathbf{b}_0, \mathbf{c}_0) = \operatorname{diag}\left\{ \frac{p'(|b_{01}|)}{2|b_{01}|}, \ldots, \frac{p'(|b_{0q}|)}{2|b_{0q}|}, \frac{p'(\|\mathbf{c}_0^{(1)}\|_{H_1})}{2\|\mathbf{c}^{(1)}\|_{H_1}} H_1, \ldots, \frac{p'(\|\mathbf{c}_0^{(p)}\|_{H_p})}{2\|\mathbf{c}^{(p)}\|_{H_p}} H_p \right\},$$

donde es claro que $\boldsymbol{\Sigma}(\mathbf{b}_0, \mathbf{c}_0) \in \mathbb{R}^{K^* \times K^*}$.

Luego, la función objetivo dada en (4.9) puede ser aproximada por

$$\frac{1}{n} \sum_{i=1}^{n} r_i^2(\mathbf{b}, \mathbf{c}) + (\mathbf{b}^{\mathrm{T}}, \mathbf{c}^{\mathrm{T}}) \boldsymbol{\Sigma}(\mathbf{b}_0, \mathbf{c}_0) (\mathbf{b}^{\mathrm{T}}, \mathbf{c}^{\mathrm{T}})^{\mathrm{T}}, \tag{4.10}$$

por lo que los estimadores obtenidos minimizando (4.10) serán cercanos a los minimizadores de (4.8) con la función de penalización dada en (4.9).

Sea $\boldsymbol{\theta} = (\mathbf{b}^{\mathrm{T}}, \mathbf{c}^{\mathrm{T}})^{\mathrm{T}}$ y $\mathbf{W}_i = (\mathbf{Z}_i^{\mathrm{T}}, \mathbf{V}_i^{\mathrm{T}})^{\mathrm{T}}$. La minimización de (4.10) puede obtenerse mediante un procedimiento iterativo ponderado.

Deducción del procedimiento iterativo

Supongamos ahora que tenemos $\boldsymbol{\theta}^{(m)} = (\mathbf{b}^{(m)\mathrm{T}}, \mathbf{c}^{(m)\mathrm{T}})^{\mathrm{T}}$, los estimadores obtenidos en el paso m. Derivando (4.10) respecto de $\boldsymbol{\theta}$, obtenemos

$$\frac{1}{n} \sum_{i=1}^{n} (Y_i - \mathbf{W}_i^{\mathrm{T}} \boldsymbol{\theta}) (-\mathbf{W}_i) + \boldsymbol{\Sigma}(\boldsymbol{\theta}^{(m)}) \boldsymbol{\theta} = 0.$$

Reescribiendo,

$$-\frac{1}{n} \sum_{i=1}^{n} \left(Y_i \mathbf{W}_i - \mathbf{W}_i \mathbf{W}_i^{\mathrm{T}} \boldsymbol{\theta} \right) + \boldsymbol{\Sigma}(\boldsymbol{\theta}^{(m)}) \boldsymbol{\theta} = 0$$

o, equivalentemente,

$$-\frac{1}{n} \sum_{i=1}^{n} Y_i \mathbf{W}_i + \left(\frac{1}{n} \sum_{i=1}^{n} \mathbf{W}_i \mathbf{W}_i^{\mathrm{T}} + \boldsymbol{\Sigma}(\boldsymbol{\theta}^{(m)}) \right) \boldsymbol{\theta} = 0\,.$$

Finalmente, de la ecuación anterior se tiene que el paso iterativo es

$$\boldsymbol{\theta}^{(m+1)} = \left(\frac{1}{n} \sum_{i=1}^{n} \mathbf{W}_i \mathbf{W}_i^{\mathrm{T}} + \boldsymbol{\Sigma}(\boldsymbol{\theta}^{(m)}) \right)^{-1} \left(\frac{1}{n} \sum_{i=1}^{n} \mathbf{W}_i Y_i \right). \tag{4.11}$$

Finalmente, obtenemos el siguiente algoritmo para minimizar (4.10):

Algoritmo iterativo ponderado

Consideremos $\epsilon > 0$ como criterio de detención, entonces:

1. Calcular un valor inicial $\boldsymbol{\theta}^{(0)}$ como el mencionado en el Paso 1 del Algoritmo 1.

2. Para $m = 0, 1, 2, \ldots$ calcular $\boldsymbol{\theta}^{(m+1)}$ como en (4.11).

3. Detener cuando la diferencia sea suficientemente pequeña, por ejemplo, si

$$\frac{\|\boldsymbol{\theta}^{(m+1)} - \boldsymbol{\theta}^{(m)}\|}{\|\boldsymbol{\theta}^{(m)}\|} < \epsilon.$$

Capítulo 5

Estudio de simulación

En este capítulo se describen los resultados de un estudio de simulación diseñado para analizar el desempeño del método propuesto de estimación y selección de variables bajo diferentes escenarios.

Para ello, las muestras $\{(Y_i, \mathbf{Z}_i^{\mathrm{T}}, \mathbf{X}_i^{\mathrm{T}})^{\mathrm{T}}\}_{i=1}^n$ son generadas con la misma distribución que $(Y, \mathbf{Z}^{\mathrm{T}}, \mathbf{X}^{\mathrm{T}})^{\mathrm{T}}$, $\mathbf{Z} = (Z_1, \ldots, Z_q)^{\mathrm{T}} \in \mathbb{R}^q$, $\mathbf{X} = (X_1, \ldots, X_p)^{\mathrm{T}} \in \mathbb{R}^p$, con $p, q \in \{2, 4, 6, 10\}$ y tres diferentes tamaños de muestra $n = 100$, 200 y 400. En todos los casos, la variable respuesta y las covariables satisfacen el modelo parcialmente lineal aditivo

$$Y = \mu + \boldsymbol{\beta}^{\mathrm{T}}\mathbf{Z} + \sum_{j=1}^{p} \eta_j(X_j) + u\,,$$

con $\mu = 0$. Las covariables $\mathbf{Z}_i = (Z_{i1}, \ldots, Z_{iq})^{\mathrm{T}}$ son generadas con distribución normal multivariada con vector de medias cero y matriz de varianza-covarianza Σ definida por $(\Sigma)_{i,\ell} = \mathrm{cov}(Z_{ik}, Z_{i\ell}) = 0.5^{|k-\ell|}$ para $1 \leq k, \ell \leq q$. Las variables $X_{ij} \sim U(0,1)$, $1 \leq j \leq p$ son independientes entre sí e independientes del vector $\mathbf{Z}_i$. El error u se distribuye como $N(0,1)$ y es independiente del vector de covariables $(\mathbf{Z}^{\mathrm{T}}, \mathbf{X}^{\mathrm{T}})$. Para garantizar identificabilidad, las funciones aditivas satisfacen que $\mathbb{E}\eta_j(X_j) = 0$ para todo $j = 1, \ldots, p$.

Teniendo en cuenta que para los diferentes valores de q y p habrá una cantidad menor o igual de componentes significativas, llamaremos q^* a la cantidad de componentes no nulas de la parte lineal y p^* a la cantidad de componentes no nulas de la parte aditiva. Claramente, $q^* \leq q$ y $p^* \leq p$.

Para seleccionar los parámetros de penalización auxiliares λ_1 y λ_2 del procedimiento adaptativo, consideramos el criterio de BIC (4.4) definido en la Sección

4.2 y la grilla $\mathcal{G} \subset \mathbb{R}^2$ definida como

$$\mathcal{G} = \{(\ell_1, \ell_2) : \ell_1 \in \Lambda_1, \ell_2 \in \Lambda_2\}, \tag{5.1}$$

donde $\Lambda_1 = \Lambda_2 = \{0, 0.01, 0.02, \ldots, 0.10, 0.15, \ldots, 0.30, 0.40, \ldots, 0.80, 1, \ldots, 2\}$. Esta grilla está estructurada de la siguiente forma: para cada parámetro, se comienza con una primera parte más fina, con paso 0.01 hasta 0.1, seguido de un paso 0.05 hasta 0.30, luego un paso 0.1 hasta el 0.80 y, finalmente, un paso 0.20 hasta el 2.

Para la aproximación de las funciones aditivas, hemos empleado splines cúbicos, lo que garantiza flexibilidad en la representación de las componentes no lineales. Dado que ninguna de las funciones consideradas presenta cambios dramáticos en sus derivadas, se han considerado nodos equiespaciados.

Para la selección del parámetro de suavizado $\mathbf{k}$, se considera el caso $k_j = k$, para $1 \leq j \leq p$ y, por ende el valor de k se selecciona utilizando el criterio de BIC (4.6) mencionado en la Sección 4.3. Teniendo en cuenta que se utilizan splines cúbicos, tal como se sugiere en He et al. (2002), el valor de k para cada escenario se elige en el conjunto $\{4, 5, \ldots, 13\}$ para $n = 100$ y $n = 200$, y $\{4, 5, \ldots, 14\}$ para $n = 400$.

Para cada uno de los tres tamaños de muestra (n = 100, 200 y 400) y para cada uno de los escenarios considerados que describiremos a continuación, el procedimiento de estimación y selección de variables se realiza sobre $N_{rep} = 500$ replicaciones. En las Secciones 5.2 y 5.3 presentaremos medidas de calidad de la selección de variables y de la estimación, respectivamente, sobre las 500 replicaciones realizadas para cada combinación de escenario y tamaño de muestra. Finalmente en la Sección 5.4 detallaremos los resultados obtenidos en el estudio de simulación.

5.1. Escenarios

En esta sección se presentan doce diferentes escenarios considerados en este estudio de simulación para diferentes valores de p, q, p^*, q^* y n. Por simplicidad, se consideró $q^* = p^*$. Además, para mayor claridad en la notación, definimos $r^* = q^* = p^*$. Para cada una de estas combinaciones de q, p y r^*, se definen, en cada escenario, subvectores del vector $(3, 1.5, 2, -1.5, 0, \ldots, 0)^{\mathrm{T}} \in \mathbb{R}^{10}$, y las funciones aditivas no nulas consideradas serán algunas de las siguientes cuatro funciones:

$$\eta_1(x) = 3(2x-1)^2 - 1, \qquad \eta_2(x) = 60x^3 - 90x^2 + 30x,$$
$$\eta_3(x) = 5x - \frac{5}{2}, \qquad \eta_4(x) = 2\sin(\pi x) - \frac{4}{\pi}.$$

Los escenarios fueron ordenados en función de dos criterios principales: primero, se prioriza el aumento de la cantidad de covariables en la parte lineal (q), y dentro de cada valor de q, se incrementa progresivamente el número de covariables en la parte aditiva (p). Este orden permite analizar cómo cambia el desempeño del estimador al incrementar la complejidad del modelo y la cantidad de variables irrelevantes en cada componente. A continuación se presentan los diferentes escenarios considerados:

i. Escenario I

- $q = 2$, $p = 2$, $r^* = 2$.
- $\boldsymbol{\beta} = (3, 1.5)^{\mathrm{T}}$.
- Las $p = 2$ funciones aditivas son

$$\eta_1(x) = 3(2x-1)^2 - 1 \qquad \eta_2(x) = 60x^3 - 90x^2 + 30x.$$

ii. Escenario II

- $q = 2$, $p = 4$, $r^* = 2$.
- $\boldsymbol{\beta} = (3, 1.5)^{\mathrm{T}}$.
- Las $p = 4$ funciones aditivas son

$$\eta_1(x) = 3(2x-1)^2 - 1 \qquad \eta_2(x) = 60x^3 - 90x^2 + 30x$$

$$\eta_j(x) \equiv 0 \text{ para } j = 3, 4.$$

iii. Escenario III

- $q = 2$, $p = 6$, $r^* = 2$.
- $\boldsymbol{\beta} = (3, 1.5)^{\mathrm{T}}$.
- Las $p = 6$ funciones aditivas son

$$\eta_1(x) = 3(2x-1)^2 - 1 \qquad \eta_2(x) = 60x^3 - 90x^2 + 30x$$

$$\eta_j(x) \equiv 0 \text{ para } j = 3, \ldots, 6.$$

iv. Escenario IV

- $q = 4$, $p = 2$, $r^* = 2$.
- $\boldsymbol{\beta} = (3, 1.5, 0, 0)^{\mathrm{T}}$.

- Las $p = 2$ funciones aditivas son

$$\eta_1(x) = 3(2x-1)^2 - 1 \qquad \eta_2(x) = 60x^3 - 90x^2 + 30x.$$

v. Escenario V

- $q = 4$, $p = 4$, $r^* = 2$.
- $\boldsymbol{\beta} = (3, 1.5, 0, 0)^{\mathrm{T}}$.
- Las $p = 2$ funciones aditivas son

$$\eta_1(x) = 3(2x-1)^2 - 1 \qquad \eta_2(x) = 60x^3 - 90x^2 + 30x$$
$$\eta_j(x) \equiv 0 \text{ para } j = 3, 4.$$

vi. Escenario VI

- $q = 4$, $p = 6$, $r^* = 4$.
- $\boldsymbol{\beta} = (3, 1.5, 2, -1.5)^{\mathrm{T}}$.
- Las $p = 6$ funciones aditivas son

$$\eta_1(x) = 3(2x-1)^2 - 1 \qquad \eta_2(x) = 60x^3 - 90x^2 + 30x$$
$$\eta_3(x) = 5x - \frac{5}{2} \qquad \eta_4(x) = 2\sin(\pi x) - \frac{4}{\pi}$$
$$\eta_j(x) \equiv 0 \text{ para } j = 5, 6.$$

vii. Escenario VII

- $q = 6$, $p = 2$, $r^* = 2$.
- $\boldsymbol{\beta} = (3, 1.5, 0, 0, 0, 0)^{\mathrm{T}}$.
- Las $p = 2$ funciones aditivas son

$$\eta_1(x) = 3(2x-1)^2 - 1 \qquad \eta_2(x) = 60x^3 - 90x^2 + 30x.$$

viii. Escenario VIII

- $q = 6$, $p = 4$, $r^* = 4$.
- $\boldsymbol{\beta} = (3, 1.5, 2, -1.5, 0, 0)^{\mathrm{T}}$.
- Las $p = 4$ funciones aditivas son

$$\eta_1(x) = 3(2x-1)^2 - 1 \qquad \eta_2(x) = 60x^3 - 90x^2 + 30x$$
$$\eta_3(x) = 5x - \frac{5}{2} \qquad \eta_4(x) = 2\sin(\pi x) - \frac{4}{\pi}.$$

ix. Escenario IX

- $q = 6$, $p = 6$, $r^* = 2$.
- $\boldsymbol{\beta} = (3, 1.5, 0, 0, 0, 0)^{\mathrm{T}}$.
- Las $p = 2$ funciones aditivas son

$$\eta_1(x) = 3(2x-1)^2 - 1 \qquad \eta_2(x) = 60x^3 - 90x^2 + 30x.$$

$$\eta_j(x) \equiv 0 \text{ para } j = 3, 4, 5, 6.$$

x. Escenario X

- $q = 6$, $p = 6$, $r^* = 4$.
- $\boldsymbol{\beta} = (3, 1.5, 0, 0, 0, 0)^{\mathrm{T}}$.
- Las $p = 6$ funciones aditivas son

$$\eta_1(x) = 3(2x-1)^2 - 1 \qquad \eta_2(x) = 60x^3 - 90x^2 + 30x.$$

$$\eta_3(x) = 5x - \frac{5}{2} \qquad \eta_4(x) = 2\sin(\pi x) - \frac{4}{\pi}$$

$$\eta_j(x) \equiv 0 \text{ para } j = 5, 6.$$

xi. Escenario XI

- $q = 10$, $p = 10$, $r^* = 2$.
- $\boldsymbol{\beta} = (3, 1.5, 0, 0, 0, 0, 0, 0, 0, 0)^{\mathrm{T}}$.
- Las $p = 10$ funciones aditivas son

$$\eta_1(x) = 3(2x-1)^2 - 1 \qquad \eta_2(x) = 60x^3 - 90x^2 + 30x$$

$$\eta_j(x) \equiv 0 \text{ para } j = 3, \ldots, 10.$$

xii. Escenario XII

- $q = 10$, $p = 10$, $r^* = 4$.
- $\boldsymbol{\beta} = (3, 1.5, 2, -1.5, 0, 0, 0, 0, 0, 0)^{\mathrm{T}}$.
- Las $p = 10$ funciones aditivas son

$$\eta_1(x) = 3(2x-1)^2 - 1 \qquad \eta_2(x) = 60x^3 - 90x^2 + 30x$$

$$\eta_3(x) = 5x - \frac{5}{2} \qquad \eta_4(x) = 2\sin(\pi x) - \frac{4}{\pi}$$

$$\eta_j(x) \equiv 0 \text{ para } j = 5, \ldots, 10.$$

El Escenario I considera el caso más simple donde $q = q^* = 2$ y $p = p^* = 2$, es decir, todas las covariables en ambas partes del modelo son significativas. Este escenario permite evaluar el estimador en una situación ideal, sin presencia de ruido o covariables irrelevantes. En los Escenarios II y III, se mantiene $q = q^* = 2$, mientras que $p > p^* = 2$, aumentando a $p = 4$ y $p = 6$, respectivamente. Esto introduce covariables irrelevantes en la parte aditiva, permitiendo analizar el impacto de estas variables en la capacidad del estimador para identificar correctamente las covariables significativas.

Los Escenarios IV, V y VI consideran $q = 4$ con diferentes combinaciones de p variando en $2, 4$ y 6, respectivamente. En el Escenario IV, $r^* = 2$, lo que lo hace análogo al Escenario II en cuanto a la presencia de dos covariables irrelevantes pero esta vez sólo en la parte lineal, permitiendo ahora evaluar cómo se desempeña el estimador cuando se introduce ruido en esta componente. En el Escenario V, $r^* = 2$ y se introducen covariables irrelevantes tanto en la parte lineal como en la aditiva dado que $q = 4$ y $p = 4$, buscando evaluar el comportamiento del estimador bajo la presencia de ruido en ambas componentes (lineal y aditiva). Finalmente, en el Escenario VI, $r^* = 4$ y, por lo tanto, todas las variables de la parte lineal son significativas, aunque $p = 6$ y, por lo tanto, el número de covariables es mayor en la parte aditiva, teniendo así dos covariables irrelevantes en esta componente. Esto permite evaluar el impacto de aumentar la dimensionalidad cuando todas las variables de la parte lineal contribuyen al modelo.

En los Escenarios VII, VIII, IX y X, se fija $q = 6$ y se analizan distintas configuraciones de p y de las variables significativas. En los Escenarios VII y VIII, $p = p^* = 2$ y $p = p^* = 4$, respectivamente, lo que permite evaluar el impacto de mantener todas las covariables aditivas como significativas mientras se incrementa p. En los Escenarios IX y X, se toma $p = 6$ y $q = 6$, pero con $r^* = 2$ en el Escenario IX y $r^* = 4$ en el Escenario X. Esto introduce tanto covariables irrelevantes en ambas partes del modelo, como un aumento en la cantidad de covariables significativas.

Finalmente, en los Escenarios XI y XII, se aumenta la dimensionalidad a $q = p = 10$. En el Escenario XI, $r^* = 2$, mientras que en el Escenario XII, $r^* = 4$. Ambos escenarios incluyen covariables irrelevantes tanto en la parte lineal como en la aditiva, pero el primero es un escenario en el que el modelo es más esparso y en el segundo se considera una mayor cantidad de variables significativas. Esto permite explorar el desempeño del estimador en un contexto de mayor dimensionalidad con diferentes grados de relevancia de las variables.

En resumen, la organización de los escenarios permite analizar de forma progresiva el impacto del aumento en la cantidad de covariables irrelevantes, así como el efecto de incrementar la dimensionalidad del modelo y la proporción de varia-

bles significativas en cada componente. Esto asegura una evaluación exhaustiva y sistemática del estimador propuesto bajo diversas condiciones de simulación.

5.2. Medidas de selección de variables

Se utilizaron diferentes medidas para evaluar el desempeño en la selección de variables, tanto para las componentes lineales como para las no paramétricas. Para cada escenario, para cada muestra (de las 500 generadas) y cada una de las partes del modelo (la lineal, la aditiva y el modelo completo), se definieron las siguientes métricas:

C: la cantidad de componentes correctamente estimadas como cero;

IC: la cantidad de componentes incorrectamente estimadas como cero; y

CF: vale 1 si fueron seleccionadas sólo las variables relevantes y vale 0 en caso contrario.

Posteriormente, se calculó el promedio de cada métrica sobre las $N_{rep} = 500$ replicaciones, obteniendo así un único valor representativo para cada una de estas tres medidas.

Dado que estas métricas también se calculan de forma independiente para cada parte del modelo, introducimos los subíndices P y NP para diferenciarlas: C_P, IC_P y CF_P corresponderán a la parte paramétrica (lineal), C_{NP}, IC_{NP} y CF_{NP} corresponderán a la parte no paramétrica (aditiva), y C, IC y CF (sin subíndices) corresponderán al modelo completo, es decir, considerando ambas partes del modelo simultáneamente. La métrica C_P correspondiente a la parte lineal puede tomar valores en $[0, q - q^*]$, la métrica C_{NP} (que corresponde a la parte aditiva) puede tomar valores en $[0, p - p^*]$, mientras que la métrica C correspondiente al modelo completo toma valores en $[0, q + p - q^* - p^*]$. Por otro lado, $IC_P \in [0, q^*]$, $IC_{NP} \in [0, p^*]$ y $IC \in [0, q^* + p^*]$, mientras que las tres métricas CF_P, CF_{NP} y CF devuelven números en el intervalo $[0, 1]$.

Para un modelo que tiene q componentes lineales de las cuales q^* son realmente importantes y p componentes no paramétricas de las cuales p^* son relevantes, el desempeño óptimo y deseado en la selección de variables de la parte lineal del modelo sería el siguiente: $C_P = q - q^*$ (es decir, todas las componentes lineales no relevantes se estiman correctamente como cero), $IC_P = 0$ (ninguna de las componentes lineales relevantes es incorrectamente estimada como cero), y $CF_P = 1$ (las q^* variables lineales relevantes son seleccionadas y las $q - q^*$ que no lo son, no son

seleccionadas). De manera análoga, en la parte no paramétrica, un desempeño óptimo se reflejaría en $\mathrm{C}_{\mathrm{NP}} = p - p^*$, $\mathrm{IC}_{\mathrm{NP}} = 0$ y $\mathrm{CF}_{\mathrm{NP}} = 1$. Por último, para el modelo completo, los valores óptimos de las métricas serían $\mathrm{C} = p + q - p^* - q^*$, $\mathrm{IC} = 0$ y $\mathrm{CF} = 1$, indicando esta última métrica que las $q^* + p^*$ variables relevantes fueron seleccionas como importantes y las $q + p - q^* - p^*$ restantes como no relevantes.

5.3. Medidas de estimación

En esta sección estableceremos medidas que nos permitirán determinar la calidad de estimación de la propuesta, tanto para la componente lineal como para la componente aditiva del modelo.

Para evaluar el desempeño de la componente paramétrica, para cada una de las 500 muestras, calculamos el Error Cuadrático Medio Generalizado (GMSE, *Generalized Mean Squared Error*), definido como

$$\mathrm{GMSE} = (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})^{\mathrm{T}} \Sigma (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})$$

donde $\Sigma \in \mathbb{R}^{q \times q}$ es la matriz de varianza-covarianza de $\mathbf{Z}$. Luego, calculamos el promedio y el desvío estándar de esta medida sobre las $N_{rep} = 500$ muestras para obtener un único valor de esta medida.

Para evaluar el desempeño del estimador $\widehat{\eta}_j$ de cada función aditiva, en cada iteración computamos el Error Cuadrático Medio (ASE, *Average Squared Error*):

$$\mathrm{ASE}(\eta_j) = \frac{1}{m} \sum_{k=1}^{m} (\widehat{\eta}_j(t_k) - \eta_j(t_k))^2$$

donde $\{t_k\}_{k=1}^m$ es una grilla de $m = 1000$ puntos equidistantes en el intervalo $[0, 1]$. Como esta medida la tenemos para cada una de las componentes aditivas, para tener una medida general de la calidad de estimación de la componente no paramétrica del modelo, para cada muestra definimos la medida Raíz Cuadrada del Error Cuadrático Medio (RASE, *Root Average Squared Error*) de la siguiente forma

$$\mathrm{RASE} = \sqrt{\sum_{j=1}^{p} \mathrm{ASE}(\eta_j)}.$$

Luego, como nuevamente tenemos una medida RASE para cada iteración, calculamos el promedio y desvío estándar sobre las $N_{rep} = 500$ muestras.

Tanto la medida asociada a la componente paramétrica como la asociada a la componente no paramétrica fueron consideradas también para los estimadores

Oracle. Los estimadores Oracle son aquellos que se obtienen considerando el modelo que utiliza únicamente las covariables significativas. En nuestro caso, sería aquel que considera únicamente las primeras q^* covariables de la parte lineal y las primeras p^* de la parte aditiva, es decir, aquel que tiene como vector de covariables de la componente lineal a $\widetilde{\mathbf{Z}} = (Z_1, \ldots, Z_{q^*})^{\mathrm{T}}$ y como vector de covariables de la componente aditiva $\widetilde{\mathbf{X}} = (X_1, \ldots, X_{p^*})^{\mathrm{T}}$. Estos estimadores son calculados por el estimador que no selecciona variables dado en Boente y Martínez (2023) utilizando como función de pérdida $\rho(t) = t^2$. Las notaciones utilizadas para los resultados de las medidas GMSE y RASE para los estimadores Oracle serán Oracle(GMSE) y Oracle(RASE), respectivamente.

5.4. Resultados

Esta sección presenta los resultados obtenidos a partir del estudio de simulación diseñado para evaluar el desempeño del procedimiento propuesto en distintos escenarios. El análisis se centra en tres aspectos principales: la capacidad del estimador para seleccionar variables de forma adecuada, la calidad de la estimación en ambas partes del modelo (paramétrica y no paramétrica), y el comportamiento en la selección automática de los parámetros involucrados en la penalización y el suavizado.

Para evaluar la selección de variables se presentan tablas que resumen en cada escenario las medidas utilizadas (C, IC, CF) para la parte lineal, para la parte aditiva, y para el modelo completo. Estas tablas permiten analizar el comportamiento global del método en términos de su capacidad para identificar correctamente las variables relevantes y aquellas irrelevantes.

En cuanto a la calidad de la estimación, se utilizaron dos medidas: GMSE, que evalúa la parte paramétrica del modelo, y RASE, que mide el error asociado a la estimación de la parte aditiva. Los valores de estas métricas se resumen tanto en tablas como mediante boxplots, lo que permite analizar el comportamiento del procedimiento en función del tamaño muestral y la complejidad del escenario. Además, se presentan boxplots funcionales que ilustran la variabilidad en la estimación de las funciones aditivas, facilitando una representación visual de su desempeño.

Respecto a los parámetros auxiliares involucrados en la penalización, se analizaron los valores seleccionados de λ_1 y λ_2, los cuales se presentan mediante boxplots agrupados por tamaño muestral. Por último, se incluyen histogramas que muestran la distribución de los valores elegidos para el parámetro de suavizado k en los diferentes escenarios simulados.

Esta sección se organiza de la siguiente manera: la Sección 5.4.1 presenta los resultados sobre la calidad de la selección de variables; la Sección 5.4.2, sobre la calidad de la estimación; la Sección 5.4.3, sobre la selección de los parámetros de penalización; y la Sección 5.4.4, sobre la selección de los parámetros de suavizado.

5.4.1. Sobre la calidad de la selección de variables

En el Cuadro 5.1 se muestra, para cada uno de los doce escenarios y los diferentes tamaños de muestra, la cantidad de componentes promedio correctamente estimadas como cero (las del tipo C), la cantidad de componentes promedio incorrectamente estimadas como cero (las del tipo IC) y la proporción de modelos en las que se seleccionaron las variables correctas (las del tipo CF), para la parte lineal por un lado, para la parte aditiva por otro y para el modelo completo.

En primer lugar destacamos que, tal como se esperaría, a medida que el tamaño de muestra aumenta, todas estas medidas mejoran, acercándose a sus valores óptimos.

En lo que respecta a la medida C_P, se observa que toma el valor 0 en los Escenarios I, II, III y VI, para los tres tamaños de muestra considerados, situación que es esperable ya que estos cuatro escenarios son aquellos en los que todas las covariables de la parte lineal son relevantes. Es de interés analizar entonces los resultados para C_P en el resto de los escenarios. Como se puede observar en estos casos, si bien el desempeño no fue perfecto, es decir, no se alcanzaron valores de C_P iguales a $q-q^*$, los resultados obtenidos fueron bastante satisfactorios, y se observa que tienden a mejorar a medida que el tamaño muestral aumenta, acercándose progresivamente al valor óptimo. En cuanto a la parte aditiva, análogamente a lo anterior, el valor de C_{NP} fue 0 en los Escenarios I, IV, VII y VIII, que coinciden con aquellos en los que todas las covariables de la parte aditiva son significativas. Es de interés entonces analizar el resto de los escenarios para esta medida. Para todos los valores de muestra considerados, nuevamente, se puede ver que C_{NP} se acerca al valor óptimo. Más aún, cuando $n = 400$, la selección fue óptima en todos los escenarios, obteniéndose $C_{NP} = p - p^*$. Aunque ya para $n = 200$, si bien la selección no fue perfecta, se lograron valores bastante cercanos al óptimo en todos los escenarios.

Analizando la medida IC para la parte lineal, en todos los escenarios y tamaños de muestra, se obtuvo $IC_P = 0$, lo que indica que no hubo casos en los que variables relevantes de la parte lineal fueran erróneamente excluidas del modelo por el estimador. Para la parte aditiva, se observó un comportamiento similar, con valores de $IC_{NP} = 0$ en casi todos los casos. Para el tamaño de muestra $n = 100$, en

Esc	q	p	r^*	n	C_P	IC_P	CF_P	C_{NP}	IC_{NP}	CF_{NP}	C	IC	CF
				100	0	0	1	0	0	1	0	0	1
I	2	2	2	200	0	0	1	0	0	1	0	0	1
				400	0	0	1	0	0	1	0	0	1
				100	0	0	1	1.904	0	0.936	1.904	0	0.936
II	2	4	2	200	0	0	1	1.996	0	0.996	1.996	0	0.996
				400	0	0	1	2	0	1	2	0	1
				100	0	0	1	3.320	0	0.806	3.320	0	0.806
III	2	6	2	200	0	0	1	3.996	0	0.996	3.996	0	0.996
				400	0	0	1	4	0	1	4	0	1
				100	1.894	0	0.902	0	0	1	1.894	0	0.902
IV	4	2	2	200	1.988	0	0.990	0	0	1	1.988	0	0.990
				400	1.996	0	0.998	0	0	1	1.996	0	0.998
				100	1.882	0	0.900	1.974	0	0.978	3.856	0	0.884
V	4	4	2	200	1.984	0	0.986	2	0	1	3.984	0	0.986
				400	1.992	0	0.996	2	0	1	3.992	0	0.996
				100	0	0	1	1.636	0.022	0.746	1.636	0.022	0.746
VI	4	6	4	200	0	0	1	1.984	0	0.986	1.984	0	0.986
				400	0	0	1	2	0	1	2	0	1
				100	3.812	0	0.866	0	0	1	3.812	0	0.866
VII	6	2	2	200	3.968	0	0.978	0	0	1	3.968	0	0.978
				400	3.998	0	0.998	0	0	1	3.998	0	0.998
				100	1.868	0	0.890	0	0	1	1.868	0	0.890
VIII	6	4	4	200	1.964	0	0.976	0	0	1	1.964	0	0.976
				400	1.992	0	0.996	0	0	1	1.992	0	0.996
				100	3.696	0	0.812	3.968	0	0.968	7.664	0	0.790
IX	6	6	2	200	3.868	0	0.954	3.998	0	0.998	7.866	0	0.954
				400	3.982	0	0.992	4	0	1	7.982	0	0.992
				100	1.764	0	0.816	1.904	0.036	0.872	3.668	0.036	0.726
X	6	6	4	200	1.928	0	0.958	1.992	0	0.992	3.920	0	0.950
				400	1.982	0	0.990	2	0	1	3.982	0	0.990
				100	6.934	0	0.604	7.940	0	0.946	14.874	0	0.582
XI	10	10	2	200	7.486	0	0.870	7.994	0	0.994	15.480	0	0.864
				400	7.876	0	0.972	8	0	1	15.876	0	0.972
				100	4.782	0	0.482	5.870	0.078	0.814	10.652	0.078	0.414
XII	10	10	4	200	5.518	0	0.842	5.984	0.004	0.982	11.502	0.004	0.836
				400	5.844	0	0.962	6	0	1	11.844	0	0.962

Cuadro 5.1: Resultados de las medidas de selección de variables: C_P, IC_P y CF_P para la componente lineal; C_{NP}, IC_{NP}, CF_{NP} para la componente no paramétrica; y C, IC y CF para el modelo completo.

los Escenarios VI, X y XII se registraron valores de IC_{NP} cercanos a cero pero no exactamente iguales. Estos escenarios tienen la particularidad de ser los escenarios con los tamaños de p más grandes ($p = 6$ y 10) pero menos esparsos en la parte

aditiva, es decir, aquellos que tienen algunas (pero pocas) componentes aditivas irrelevantes. Para $n = 200$, en 11 de los 12 escenarios se obtuvo $\mathrm{IC}_{\mathrm{NP}} = 0$, con la única excepción del Escenario XII, donde se registró un valor de 0.004. Este escenario corresponde al modelo con dimensión más grande en la parte aditiva ($p = 10$) con 4 variables relevantes en la parte aditiva.

En cuanto a la medida CF, en la parte lineal se observa una selección óptima en los Escenarios I, II, III y VI para todos los tamaños de muestra, alcanzándose $\mathrm{CF}_{\mathrm{P}} = 1$. Estos escenarios corresponden, como se mencionó en el análisis de C_{P}, a los casos en los que todas las covariables de la parte lineal son significativas. En los demás escenarios, aunque la selección no fue perfecta, los valores de CF_{P} se mantuvieron cercanos a 1 para todos los tamaños de muestra, a excepción de los Escenarios XI y XII en los que para el tamaño de muestra más pequeño alcanzó una proporción de 0.604 y 0.482 de submodelos (mirando exclusivamente la parte lineal) correctos. En lo que respecta a la parte aditiva, el estimador identificó perfectamente qué variables eran relevantes y cuáles no para esta componente en los Escenarios I, IV, VII y VIII para todos los tamaños de muestra, en concordancia con los resultados observados en C_{NP} y para algunas combinaciones de escenarios y tamaños de muestra particulares. Además, para $n = 200$ y $n = 400$, la selección de las componentes aditivas fue muy precisa en todos los escenarios, con valores de $\mathrm{CF}_{\mathrm{NP}}$ superiores a 0.982 en $n = 200$ y selección perfecta en $n = 400$. Para $n = 100$, el desempeño fue sumamente satisfactorio en la mayoría de los escenarios, con un rendimiento levemente menor en los Escenarios III, VI y XII, donde se registraron valores de $\mathrm{CF}_{\mathrm{NP}}$ de 0.806, 0.746 y 0.814, respectivamente. Finalmente, analizando la medida CF para el modelo completo, los valores obtenidos son bastante cercanos al óptimo, especialmente para los casos de tamaños de muestra más grande. Particularmente, para $n = 400$, en los Escenarios I, II, III y VI se obtuvo una selección perfecta, mientras que en los demás escenarios el desempeño fue cercano al óptimo, con un valor mínimo de 0.962 en el Escenario XII. Para $n = 200$, se observó una disminución progresiva de los valores de CF desde 1 hasta 0.836 en el peor de los casos y que se corresponde con el Escenario XII, lo que refleja un muy buen desempeño del método.

En resumen, los resultados muestran que el procedimiento propuesto logra una adecuada selección de variables, con un desempeño que mejora, como es de esperarse, al aumentar el tamaño de la muestra, y que es especialmente sólido en modelos de menor dimensión. Para modelos más grandes y con menor nivel de esparsidad, la selección se vuelve más desafiante, pero el estimador sigue mostrándose útil en la mayoría de los escenarios considerados.

5.4.2. Sobre la calidad de la estimación

En los Cuadros 5.2 y 5.3 se presentan, para cada escenario y cada tamaño de muestra, los valores obtenidos de GMSE y RASE promedios para los estimadores penalizados propuestos en este trabajo de tesis junto con los correspondientes a los estimadores Oracle, denotados como Oracle(GMSE) y Oracle(RASE), respectivamente. Además, entre paréntesis, se indica el desvío estándar de los $N_{rep} = 500$ valores de GMSE y RASE obtenidos. Recordemos que la medida GMSE evalúa el desempeño de los estimadores en la parte lineal, mientras que el RASE mide la calidad de la estimación en la parte aditiva. En consecuencia, la buena capacidad de estimación se observará en la obtención de valores de GMSE y RASE pequeños y, para la propuesta de estimación penalizada, en la obtención de valores cercanos a los producidos por los estimadores Oracle.

En primer lugar, al examinar ambos cuadros, se observa que, como era de esperar, a medida que el tamaño de muestra aumenta, los valores de GMSE y RASE tienden a disminuir, acercándose progresivamente a Oracle(GMSE) y Oracle(RASE), respectivamente, como es deseable. Al analizar el comportamiento de los desvíos estándar de los GMSE y RASE en todos los escenarios, puede notarse que estos también disminuyen de manera sistemática al crecer el tamaño muestral.

Focalizándonos en la parte lineal, se puede observar en primer lugar que en el Escenario I, que es aquel con $q = 2$, $p = 2$ y $r^* = 2$, donde no hay variables irrelevantes, el estimador penalizado reproduce exactamente el comportamiento del estimador Oracle, tanto en términos del GMSE como de su variabilidad, independientemente del tamaño muestral. Esto sugiere que, en situaciones ideales donde no hay variables irrelevantes presentes, el procedimiento propuesto no pierde capacidad en la predicción. En el Escenario II, se mantienen fijos los valores de q y r^* pero se toma $p = 4$, los valores de GMSE del estimador penalizado y del Oracle se mantienen muy cercanos, llegando incluso para $n = 100$ a tener variabilidad levemente menor, y para $n = 200$ valor promedio de GMSE levemente menor en el estimador penalizado. En el caso del Escenario III, donde $p = 6$, para $n = 100$ el valor computado de GMSE del estimador propuesto es aproximadamente un 50 % mayor al del Oracle, aumentando incluso a más del doble el desvío. Este resultado pone de manifiesto que, cuando el número de variables irrelevantes se incrementa en relación con la cantidad de observaciones disponibles, el procedimiento se ve afectado, reflejando dificultad en realizar una predicción precisa en muestras pequeñas. Sin embargo, como es de esperarse, esta diferencia se atenúa rápidamente al aumentar el tamaño muestral, ya que para $n = 200$ y $n = 400$ el GMSE del estimador penalizado vuelve a coincidir con el del Oracle, lo que sugiere que la presencia de variables irrelevantes no constituye un obstáculo significativo cuando

Esc	q	p	r^*	n	Oracle(GMSE)	GMSE
I	2	2	2	100	0.0215 (0.0208)	0.0215 (0.0208)
				200	0.0113 (0.0117)	0.0113 (0.0117)
				400	0.0052 (0.0051)	0.0052 (0.0051)
II	2	4	2	100	0.0215 (0.0208)	0.0215 (0.0206)
				200	0.0113 (0.0117)	0.0112 (0.0117)
				400	0.0052 (0.0051)	0.0052 (0.0051)
III	2	6	2	100	0.0215 (0.0208)	0.0311 (0.0455)
				200	0.0113 (0.0117)	0.0112 (0.0116)
				400	0.0052 (0.0051)	0.0052 (0.0051)
IV	4	2	2	100	0.0215 (0.0208)	0.0280 (0.0338)
				200	0.0113 (0.0117)	0.0124 (0.0132)
				400	0.0052 (0.0051)	0.0054 (0.0055)
V	4	4	2	100	0.0215 (0.0208)	0.0274 (0.0323)
				200	0.0113 (0.0117)	0.0123 (0.0131)
				400	0.0052 (0.0051)	0.0054 (0.0054)
VI	4	6	4	100	0.0526 (0.0419)	0.0660 (0.0668)
				200	0.0228 (0.0155)	0.0230 (0.0157)
				400	0.0104 (0.0070)	0.0104 (0.0069)
VII	6	2	2	100	0.0215 (0.0208)	0.0309 (0.0393)
				200	0.0113 (0.0117)	0.0129 (0.0150)
				400	0.0052 (0.0051)	0.0054 (0.0054)
VIII	6	4	4	100	0.0526 (0.0419)	0.0592 (0.0469)
				200	0.0228 (0.0155)	0.0248 (0.0185)
				400	0.0104 (0.0070)	0.0108 (0.0073)
IX	6	6	2	100	0.0215 (0.0208)	0.0322 (0.0393)
				200	0.0113 (0.0117)	0.0134 (0.0156)
				400	0.0052 (0.0051)	0.0057 (0.0058)
X	6	6	4	100	0.0526 (0.0419)	0.0625 (0.0516)
				200	0.0228 (0.0155)	0.0250 (0.0186)
				400	0.0104 (0.0070)	0.0109 (0.0074)
XI	10	10	2	100	0.0215 (0.0208)	0.0428 (0.0531)
				200	0.0113 (0.0117)	0.0140 (0.0169)
				400	0.0052 (0.0051)	0.0059 (0.0061)
XII	10	10	4	100	0.0526 (0.0419)	0.0868 (0.0720)
				200	0.0228 (0.0155)	0.0284 (0.0217)
				400	0.0104 (0.0070)	0.0111 (0.0079)

Cuadro 5.2: Valores de GMSE promedios obtenidos para cada escenario, tanto del estimador propuesto como del estimador Oracle. Además, se reportan, entre paréntesis, los desvíos estándar de los GMSE obtenidos para cada muestra.

la muestra es suficientemente grande. Un patrón similar se observa al comparar los Escenarios IV, V y VI, todos con $q = 4$, variando p progresivamente en 2, 4 y 6, pero con diferentes valores de r^*. En los Escenarios IV y V, con $r^* = 2$, el

Esc	q	p	r^*	n	Oracle(RASE)	RASE
I	2	2	2	100	0.3127 (0.1169)	0.3127 (0.1169)
				200	0.1953 (0.0674)	0.1953 (0.0674)
				400	0.1289 (0.0454)	0.1289 (0.0454)
II	2	4	2	100	0.3127 (0.1169)	0.3121 (0.1656)
				200	0.1953 (0.0674)	0.1836 (0.0576)
				400	0.1289 (0.0454)	0.1228 (0.0387)
III	2	6	2	100	0.3127 (0.1169)	0.5246 (0.5372)
				200	0.1953 (0.0674)	0.1821 (0.0549)
				400	0.1289 (0.0454)	0.1224 (0.0382)
IV	4	2	2	100	0.3127 (0.1169)	0.3006 (0.1000)
				200	0.1953 (0.0674)	0.1919 (0.0628)
				400	0.1289 (0.0454)	0.1277 (0.0433)
V	4	4	2	100	0.3127 (0.1169)	0.2847 (0.0986)
				200	0.1953 (0.0674)	0.1830 (0.0564)
				400	0.1289 (0.0454)	0.1228 (0.0388)
VI	4	6	4	100	0.4590 (0.1964)	0.6099 (0.4416)
				200	0.2766 (0.0640)	0.2752 (0.0741)
				400	0.1828 (0.0428)	0.1812 (0.0400)
VII	6	2	2	100	0.3127 (0.1169)	0.3004 (0.0985)
				200	0.1953 (0.0674)	0.1915 (0.0621)
				400	0.1289 (0.0454)	0.1274 (0.0429)
VIII	6	4	4	100	0.4590 (0.1964)	0.4160 (0.1054)
				200	0.2766 (0.0640)	0.2690 (0.0620)
				400	0.1828 (0.0428)	0.1833 (0.0416)
IX	6	6	2	100	0.3127 (0.1169)	0.2947 (0.1217)
				200	0.1953 (0.0674)	0.1821 (0.0549)
				400	0.1289 (0.0454)	0.1224 (0.0381)
X	6	6	4	100	0.4590 (0.1964)	0.4493 (0.1419)
				200	0.2766 (0.0640)	0.2725 (0.0715)
				400	0.1828 (0.0428)	0.1814 (0.0406)
XI	10	10	2	100	0.3127 (0.1169)	0.3203 (0.1496)
				200	0.1953 (0.0674)	0.1829 (0.0565)
				400	0.1289 (0.0454)	0.1225 (0.0381)
XII	10	10	4	100	0.4590 (0.1964)	0.5236 (0.1793)
				200	0.2766 (0.0640)	0.2945 (0.0968)
				400	0.1828 (0.0428)	0.1816 (0.0409)

Cuadro 5.3: Valores de RASE promedios obtenidos para cada escenario, tanto del estimador propuesto como del estimador Oracle. Además, se reportan, entre paréntesis, los desvíos estándar de los RASE obtenidos para cada muestra.

aumento del GMSE respecto del Oracle es leve, especialmente para $n = 100$, donde los valores promedian 0.0280 y 0.0274, aproximadamente un 30 % más, frente a Oracle(GMSE) = 0.0215. La diferencia disminuye a medida crece el tamaño mues-

tral, y se vuelve prácticamente despreciable para $n = 400$. En el Escenario VI, con $r^* = 4$, se evidencia una diferencia similar en los valores de GMSE con $n = 100$, alcanzando 0.0660, aproximadamente un 25 % más frente a 0.0526 del Oracle. Aún así, al pasar a muestras más grandes, estas diferencias vuelven a reducirse. En los Escenarios VII y IX, donde $q = 6$, $r^* = 2$, pero el número de variables en la parte aditiva cambia de 2 a 6, se observa para $n = 100$ una gran pérdida en la capacidad predictiva al tener valores de GMSE en el estimador penalizado de 0.0309 y 0.0322, frente a 0.0215 en el estimador Oracle. Esta situación se agrava al notar que los desvíos casi se duplican, pasando de 0.0208 en el Oracle a 0.0393 en el estimador penalizado. En el caso $n = 200$ se ve reflejado una menor pérdida en la estimación donde los valores de GMSE son 0.0129 y 0.0134, levemente mayores al 0.0113 obtenido en el Oracle. Finalmente, cuando $n = 400$, nuevamente los valores promedio y los desvíos del GMSE penalizado y Oracle son muy cercanos. En los Escenarios VIII y X, donde nuevamente $q = 6$, pero ahora $r^* = 4$ y p toma valores 4 o 6, las diferencias se apaciguan respecto a los dos escenarios anteriores, ya que por ejemplo, para $n = 100$, los GMSE pasar de 0.0592 a 0.0526 en el primer caso, y de 0.0625 a 0.0526 en el segundo, evidenciando una baja diferencia. Al aumentar los valores de n a 200 y 400, nuevamente las diferencias se reducen significativamente. En los Escenarios XI y XII, donde $q = p = 10$ y r^* toma valores 2 y 4 respectivamente, los cuales son los más complejos que hemos analizado, son particularmente reveladores. En ambos casos, los valores de GMSE para el estimador penalizado con $n = 100$ se alejan de los del Oracle de forma más notoria que en los escenarios anteriores, alcanzando 0.0428 y 0.0868 respectivamente, frente a 0.0215 y 0.0526 del Oracle. Esta diferencia pone de manifiesto que, en situaciones de mayor dimensión con tamaños de muestras pequeños, la capacidad de estimación se vuelve considerablemente más desafiante. No obstante, el patrón observado previamente vuelve a repetirse: al incrementar el tamaño muestral a 200 y luego a 400, las diferencias se vuelven menores, lo que confirma que el estimador penalizado mantiene una capacidad de ajuste razonable incluso en escenarios complejos, siempre que se disponga de un número de observaciones adecuado.

Con respecto a la parte aditiva, en el Cuadro 5.3 se observa que, a lo largo de los distintos escenarios y tamaños de muestra, al igual que lo que pasaba al analizar la componente lineal del modelo con el Cuadro 5.2, los valores de RASE promedio arrojados por el estimador penalizado son sumamente parecidos, iguales y hasta en algunos casos levemente menores, que los obtenidos por el estimador Oracle, especialmente para los tamaños de muestra más grandes, y con desvíos estándar también similares. En los Escenarios I, II y III, donde $q = 2$ y $r^* = 2$, se observa que el estimador penalizado logra un desempeño prácticamente idéntico al Oracle cuando $p = 2$, y muestra una leve mejora respecto a la del Oracle cuando $p = 4$ o $p = 6$, especialmente en tamaños muestrales mayores. Esta pequeña mejora en

los valores promedios que ocurre en algunas situaciones puede atribuirse al error de Monte Carlo asociado a la generación aleatoria de las muestras. Sin embargo, en el Escenario III y cuando la muestra es pequeña ($n = 100$), el valor promedio de RASE para el estimador penalizado se incrementa aproximadamente en un 67 % respecto del Oracle, mientras que los desvíos casi se quintuplican, pasando de 0.1169 a 0.5372. Recordemos que este escenario, con $n = 100$ ya ha dado no tan buenos resultados no sólo en la medida GMSE, sino también en lo que respecta a las medidas de calidad de selección de variables para la parte aditiva del modelo. Cuando se incrementa el número total de variables a $q = 4$ (Escenarios IV, V y VI), se mantiene la buena performance del penalizado. En los Escenarios IV y V, donde $r^* = 2$, tanto el valor promedio de RASE como el desvío del estimador propuesto son levemente inferiores a los del estimador Oracle. En el Escenario VI, donde $r^* = 4$, se evidencia para el tamaño de muestra $n = 100$, una considerable baja en la precisión de la estimación del estimador penalizado, incrementando un 33 % el valor del RASE obtenido por el estimador Oracle. Además, en este caso, se refleja un aumento considerable en la variabilidad del estimador propuesto, ya que mientras que el desvío del Oracle es de 0.1964, el desvío obtenido por el estimador penalizado es de 0.4416. Sin embargo, al aumentar el tamaño de muestra a $n = 200$ y $n = 400$, los valores de RASE disminuyen, nuevamente asimilándose a los valores obtenidos por el estimador Oracle. Si consideramos los casos en que $q = 6$, en los Escenarios VII, VIII, IX y X se observa un comportamiento del estimador penalizado similar al descripto anteriormente ya que, en casi todos los casos e independientemente del tamaño muestral, los valores de RASE promedios obtenidos para la propuesta de esta tesis son similares, y en algunos casos levemente inferiores, a los obtenidos por el estimador Oracle. Finalmente, en los escenarios más exigentes (XI y XII), donde $q = p = 10$ y se consideran $r^* = 2$, $r^* = 4$, respectivamente, el método penalizado sigue mostrando un rendimiento sólido. Por ejemplo, para $n = 400$, el RASE del penalizado es levemente mayor al del Oracle (0.1225 frente a 0.1289 para $r^* = 2$ y 0.1816 frente a 0.1828 para $r^* = 4$), lo que refuerza su capacidad para recuperar la estructura aditiva aún en situaciones de alta dimensión y mayor número de variables irrelevantes.

Con el objetivo de analizar visualmente la distribución de las métricas GMSE y RASE obtenidas en cada escenario, se incluyen seis figuras que presentan boxplots ajustados, que son boxplots adaptados a datos asimétricos (propuesto por Hubert y Vandervieren, 2008), para cada combinación de escenario y tamaño muestral. Estas representaciones gráficas permiten observar no sólo las diferencias centrales entre el estimador penalizado propuesto y el estimador Oracle, sino también el grado de dispersión de los resultados y la presencia de posibles observaciones extremas en los valores de GMSE y RASE. En particular, las Figuras 5.1, 5.2 y 5.3 muestran los boxplots ajustados correspondientes a los valores de GMSE

obtenidos en los doce escenarios de simulación. Las Figuras 5.4, 5.5 y 5.6 presentan los resultados análogos para la métrica RASE. En todos los casos, los boxplots fueron construidos a partir de los 500 valores obtenidos para cada combinación de estimador y escenario. Cada figura organiza los escenarios en filas (cuatro escenarios por figura), mientras que las columnas corresponden a los tres tamaños muestrales considerados: $n = 100$, $n = 200$ y $n = 400$. Dentro de cada celda se ubican dos boxplots paralelos, uno para el estimador Oracle (en color celeste) y otro para el estimador penalizado (en verde), lo que facilita la comparación directa entre ambos métodos.

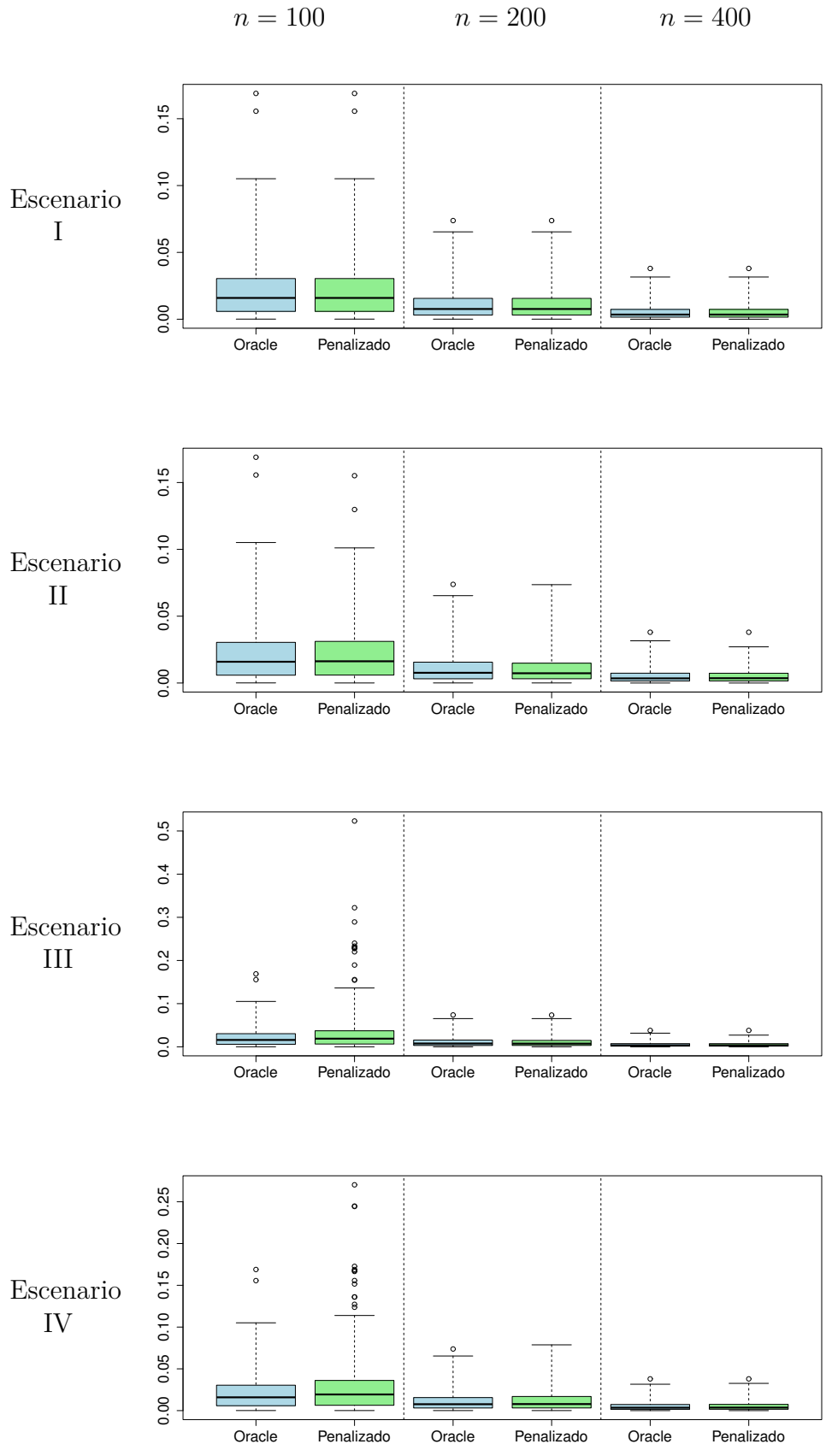

Figura 5.1: Boxplots asimétricos paralelos de la medida GMSE para el estimador Oracle y para el penalizado, para los Escenarios I a IV y los tres tamaños de muestra.

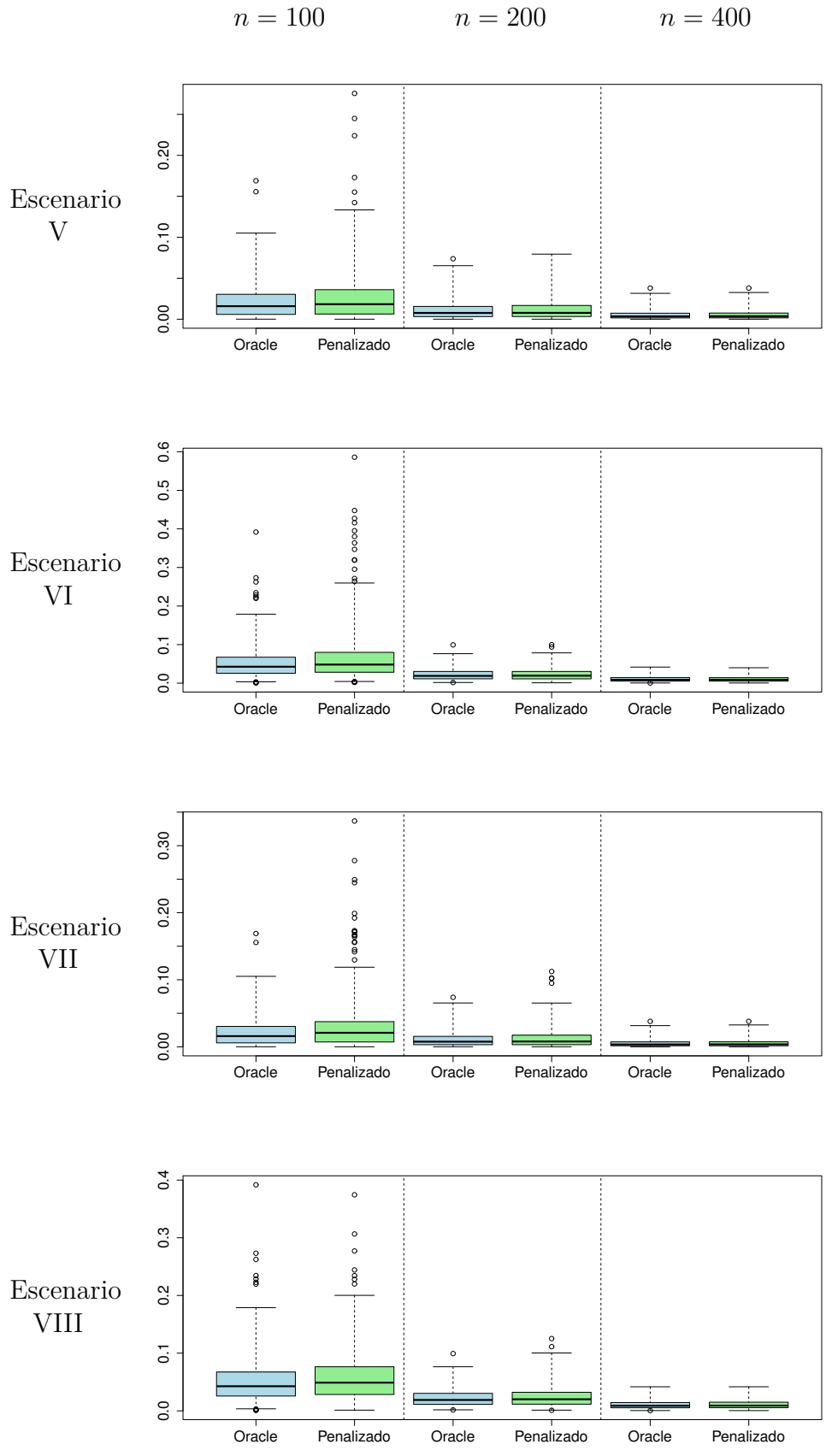

Figura 5.2: Boxplots asimétricos paralelos de la medida GMSE para el estimador Oracle y para el penalizado, para los Escenarios V a VIII y los tres tamaños de muestra.

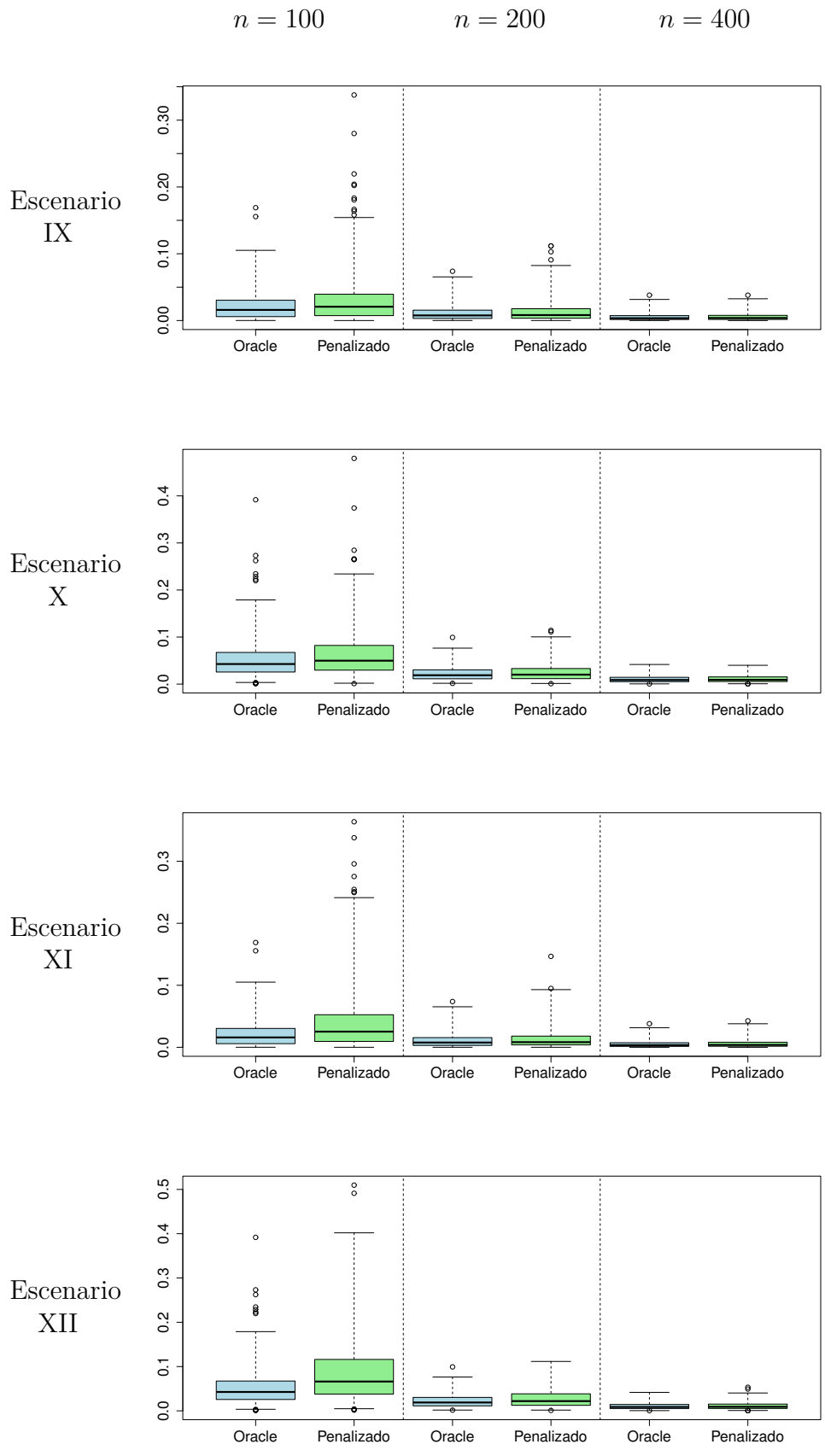

Figura 5.3: Boxplots asimétricos paralelos de la medida GMSE para el estimador Oracle y para el penalizado, para los Escenarios IX a XII y los tres tamaños de muestra.

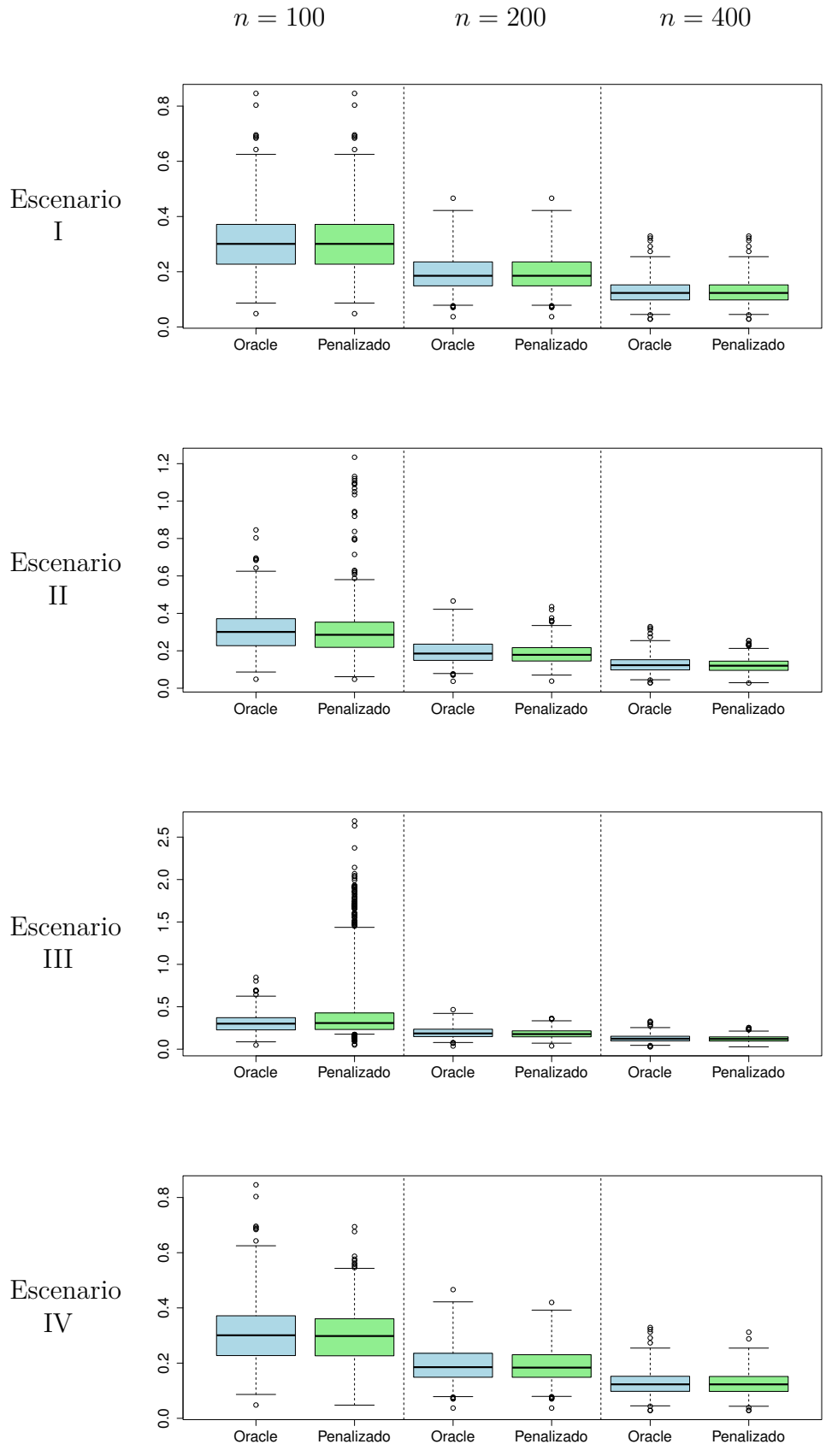

Figura 5.4: Boxplots asimétricos paralelos de la medida RASE para el estimador Oracle y para el penalizado, para los Escenarios I a IV y los tres tamaños de muestra.

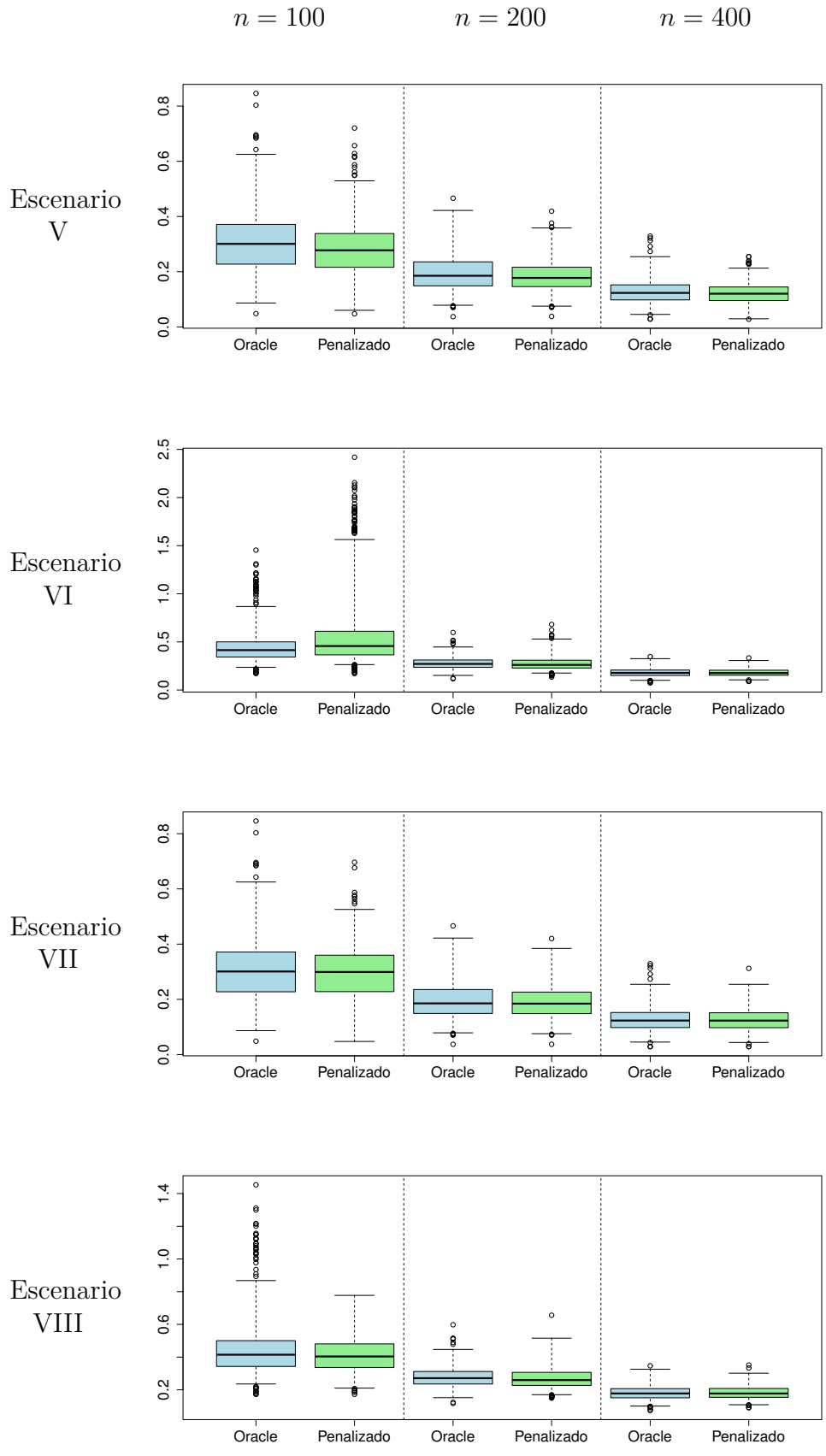

Figura 5.5: Boxplots asimétricos paralelos de la medida RASE para el estimador Oracle y para el penalizado, para los Escenarios V a VIII y los tres tamaños de muestra.

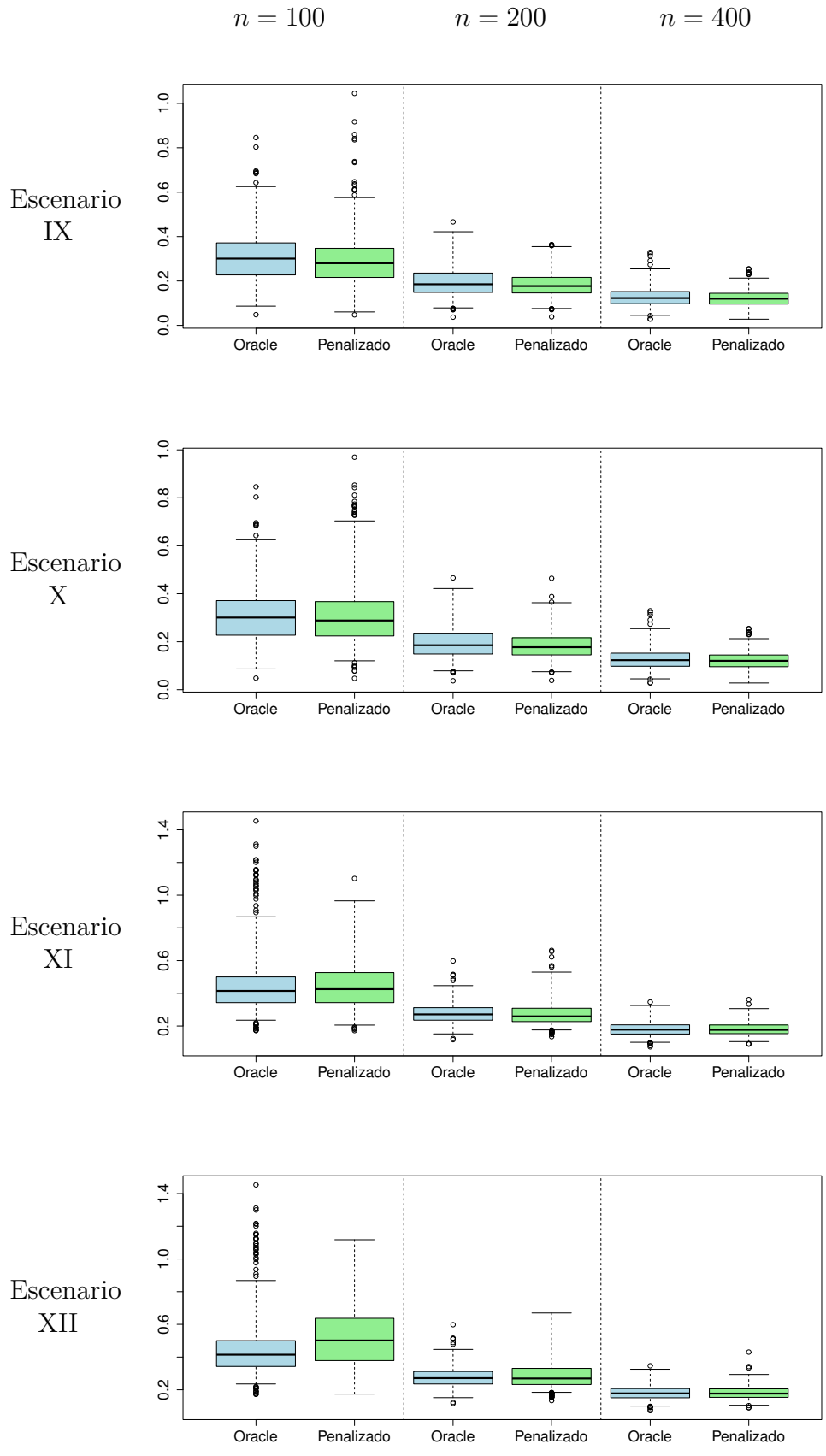

Figura 5.6: Boxplots asimétricos paralelos de la medida RASE para el estimador Oracle y para el penalizado, para los Escenarios IX a XII y los tres tamaños de muestra.

En términos generales, los resultados obtenidos a partir de los boxplots presentados en las Figuras 5.1, 5.2 y 5.3, correspondientes a los valores de GMSE, muestran un desempeño muy similar entre el estimador penalizado propuesto y el estimador Oracle. En todos los escenarios considerados y para los tres tamaños muestrales ($n = 100, 200, 400$), las cajas de ambos estimadores se encuentran a alturas comparables, sin diferencias sistemáticas a favor de uno u otro, lo que sugiere que no hay evidencia de un sesgo marcado introducido por la penalización. Las pequeñas diferencias son más notorias en algunos escenarios con el tamaño de muestra más pequeño ($n = 100$), donde se puede observar que la caja del boxplot del estimador propuesto es levemente más grande, la cola derecha es algo más larga y, en algunas oportunidades, se detectan más observaciones atípicas. En general, esta similitud en las distribuciones de los valores de GMSE indicaría que el estimador penalizado logra un ajuste muy próximo al que se obtendría si se conociera de antemano la verdadera estructura del modelo. Estos gráficos nos permiten observar, a lo largo de todos los escenarios, la disminución progresiva de la variabilidad de ambos estimadores a medida que se incrementa el tamaño muestral. Esto se refleja en una reducción visible en la amplitud de las cajas y en una menor cantidad de valores atípicos, lo que evidencia una ganancia en precisión.

Analizando en mayor detalle cada conjunto de escenarios para la medida GMSE, en la Figura 5.1, correspondiente a los Escenarios I a IV, se destaca que en el Escenario I los boxplots del estimador Oracle y del penalizado son prácticamente idénticos para todos los valores de n, lo que confirma que, en contextos ideales, el estimador propuesto reproduce de gran manera el comportamiento del estimador Oracle. En el Escenario II también se aprecia una gran similitud entre ambos métodos. Los Escenarios III y IV presentan un mayor desafío. Para el caso $n = 100$, el estimador penalizado muestra una mayor cantidad de valores atípicos respecto al Oracle, lo cual incrementa su variabilidad. No obstante, las medianas siguen siendo comparables, no tan así las amplitudes de las cajas. Esta observación es coherente con los resultados numéricos reportados en el Cuadro 5.2, donde, por ejemplo, en el Escenario III, los desvíos estándar del GMSE para el Oracle y el penalizado son 0.0208 y 0.0455, respectivamente, y en el Escenario IV, 0.0208 y 0.0338. Cabe señalar que estas diferencias se atenúan considerablemente al aumentar el tamaño de muestra, y que para $n = 400$ prácticamente desaparecen los valores atípicos en ambos casos. La Figura 5.2, que incluye los Escenarios V a VIII, exhibe un comportamiento similar. Para $n = 100$, el estimador penalizado presenta mayor dispersión, evidenciándose en una caja más ancha y colas más largas, y una cantidad mayor de valores atípicos, especialmente en los Escenarios V, VI y VII. Sin embargo, al igual que en los casos anteriores, estas diferencias se reducen notoriamente con muestras más grandes. En el Escenario VIII, en cambio, los boxplots son muy similares en los tres tamaños muestrales, mostrando un buen desempeño

del estimador penalizado incluso en contextos más complejos. Finalmente, en la Figura 5.3, correspondiente a los Escenarios IX a XII, se evalúan los casos más complejos con mayor cantidad de variables ($q = p = 6$ en los Escenarios IX y X, y $q = p = 10$ en los Escenarios XI y XII). Para $n = 100$, el estimador penalizado muestra nuevamente una mediana ubicada en una posición levemente mayor, mayor dispersión y más valores atípicos que el estimador Oracle, reflejando una mayor sensibilidad ante escenarios de alta dimensión. No obstante, al igual que en los escenarios previos, el aumento del tamaño muestral produce una mejora clara en el desempeño de ambos estimadores, disminuyendo así las diferencias entre los gráficos. Más aún, para $n = 400$ los boxplots son prácticamente idénticos entre ambos estimadores, lo que respalda la solidez de la propuesta aún en escenarios con gran cantidad de variables en la parte lineal y aditiva.

Los gráficos correspondientes a la medida RASE, presentados en las Figuras 5.4, 5.5 y 5.6, reflejan una tendencia general similar a la observada previamente para la medida GMSE. En todos los escenarios evaluados, se aprecia que al aumentar el tamaño muestral, los boxplots de los valores de RASE se achican de forma notable, indicando una mejora en los ajustes de la componente aditiva obtenidos por ambos estimadores. En particular, las medianas disminuyen con el incremento de n, acercándose a cero. Para el tamaño de muestra $n = 100$, para ambos estimadores se observan los boxplots que tienen mayor variabilidad y presencia de una cantidad considerable de datos atípicos. En algunos casos, los boxplots de la medida RASE obtenidos por el estimador Oracle son más alargados que los obtenidos por el estimador penalizado, presentando, por ejemplo en el Escenario IV, una cola derecha más larga y una mayor cantidad de datos atípicos. Estas diferencias pueden deberse a que, en muestras pequeñas, los estimadores pueden verse afectados por la mayor inestabilidad inherente al problema de estimación y a su vez a la gran estabilidad del algoritmo que computa el estimador penalizado. Al aumentar el tamaño muestral a $n = 200$, se advierte una disminución significativa de los valores atípicos, acompañada por un estrechamiento de las cajas, lo que indica un incremento en la precisión de los estimadores. Esta mejora se acentúa aún más al considerar $n = 400$, donde en todos los escenarios los boxplots presentan una forma más estrecha y simétrica, con pocos datos atípicos. Además, las medianas descienden sistemáticamente con el aumento de n, evidenciando una mejora en el ajuste.

En conclusión, los resultados obtenidos a partir de ambas métricas para evaluar la capacidad de predicción (GMSE para la parte lineal y RASE para la parte aditiva) permiten extraer conclusiones generales sobre el comportamiento del estimador penalizado propuesto. En particular, se verifica una mejora sistemática en la calidad de estimación conforme aumenta el tamaño muestral, con reducciones notables en los valores promedio de las métricas a lo largo de los distintos esce-

narios. Esta tendencia es consistente tanto en escenarios simples como en aquellos de mayor complejidad estructural. Además, se observa que la variabilidad de las métricas de evaluación es baja y comparable con la del estimador Oracle, lo cual sugiere una alta estabilidad del estimador penalizado frente a los diferentes escenarios. El buen desempeño relativo frente al estimador Oracle, incluso en situaciones con alta dimensión o con presencia de muchas variables irrelevantes, confirma la eficacia del método para recuperar adecuadamente tanto la estructura lineal como la aditiva, combinando precisión y parsimonia en la estimación.

Por otro lado, en las Figuras 5.7 a 5.17, se presentan los boxplots funcionales correspondientes a los doce escenarios analizados. Recordemos que un boxplot funcional, propuesto por Sun y Genton (2011), es una herramienta gráfica diseñada para representar la distribución de datos funcionales, es decir, aquellos que se describen mediante funciones en lugar de puntos discretos. Se adapta el concepto tradicional de boxplot, que muestra la mediana, cuartiles y valores atípicos de un conjunto de datos, para representar las funciones. En lugar de trabajar con puntos únicos, el boxplot funcional muestra la distribución de las funciones, permitiendo visualizar la variabilidad y tendencia central de los datos. En el gráfico, la caja se representa como intervalos entre los cuartiles funcionales y los valores atípicos son las funciones que se encuentran por fuera de un rango determinado.

En cada figura, se visualizan las estimaciones de las funciones aditivas para tres tamaños de muestra diferentes, $n = 100, 200, 400$. La curva verde representa la función verdadera. La disposición de las figuras sigue un esquema en el que las filas corresponden a los distintos tamaños de muestra, mientras que las columnas agrupan las estimaciones de las funciones aditivas en cada escenario. Dentro de cada figura, los límites del eje y se mantienen constantes para los tres tamaños de muestra en cada función aditiva, aunque pueden variar entre las diferentes funciones. Esta estructura permite observar con claridad cómo la estimación de las funciones aditivas mejora a medida que aumenta el tamaño de muestra, un comportamiento que coincide con lo previamente discutido. Las figuras se encuentran organizadas según el orden de los escenarios.

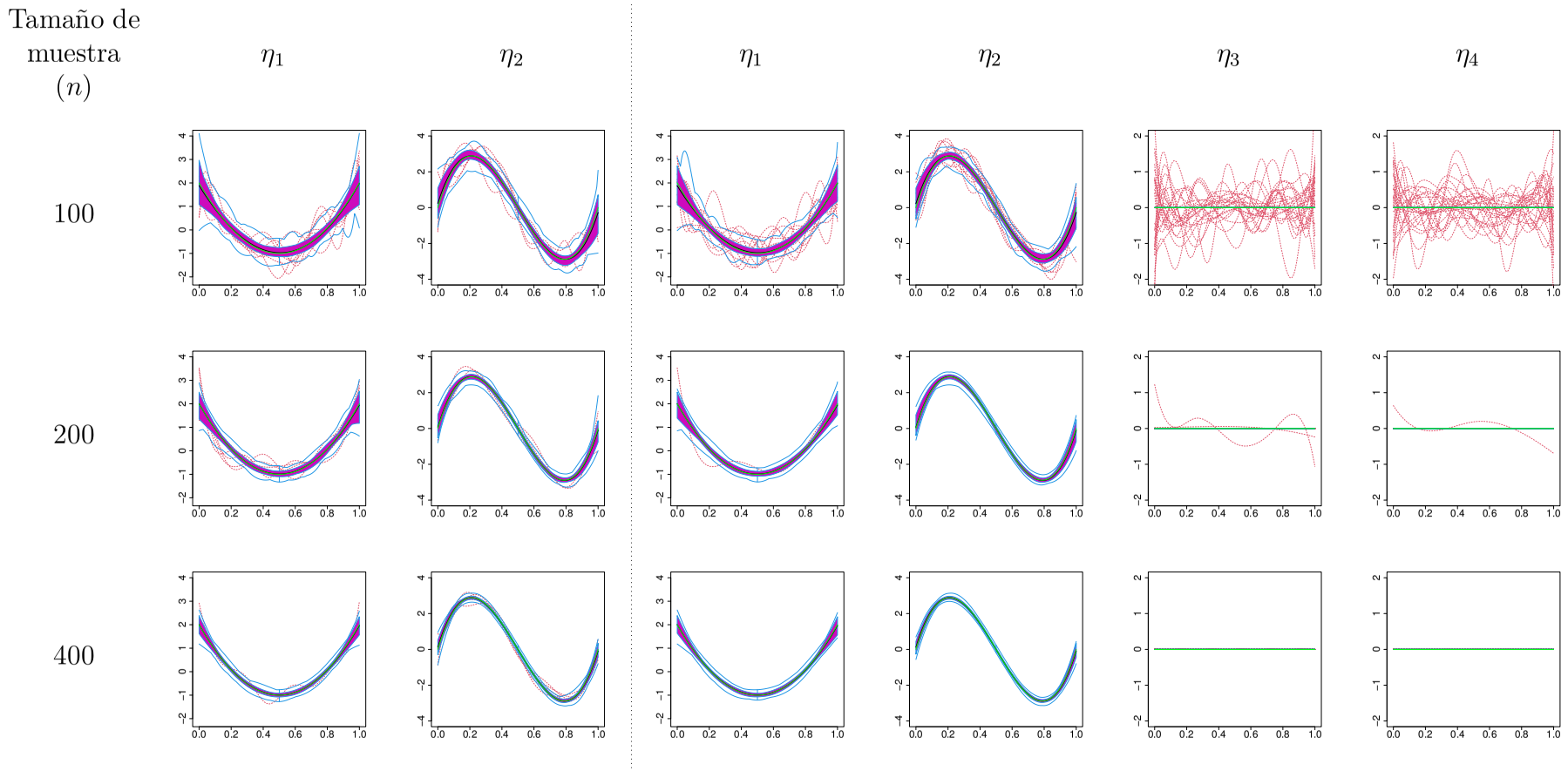

Figura 5.7: Boxplots funcionales: las dos primeras columnas corresponden al Escenario I (η_1, η_2) y las últimas cuatro columnas corresponden al Escenario II (η_1 a η_4).

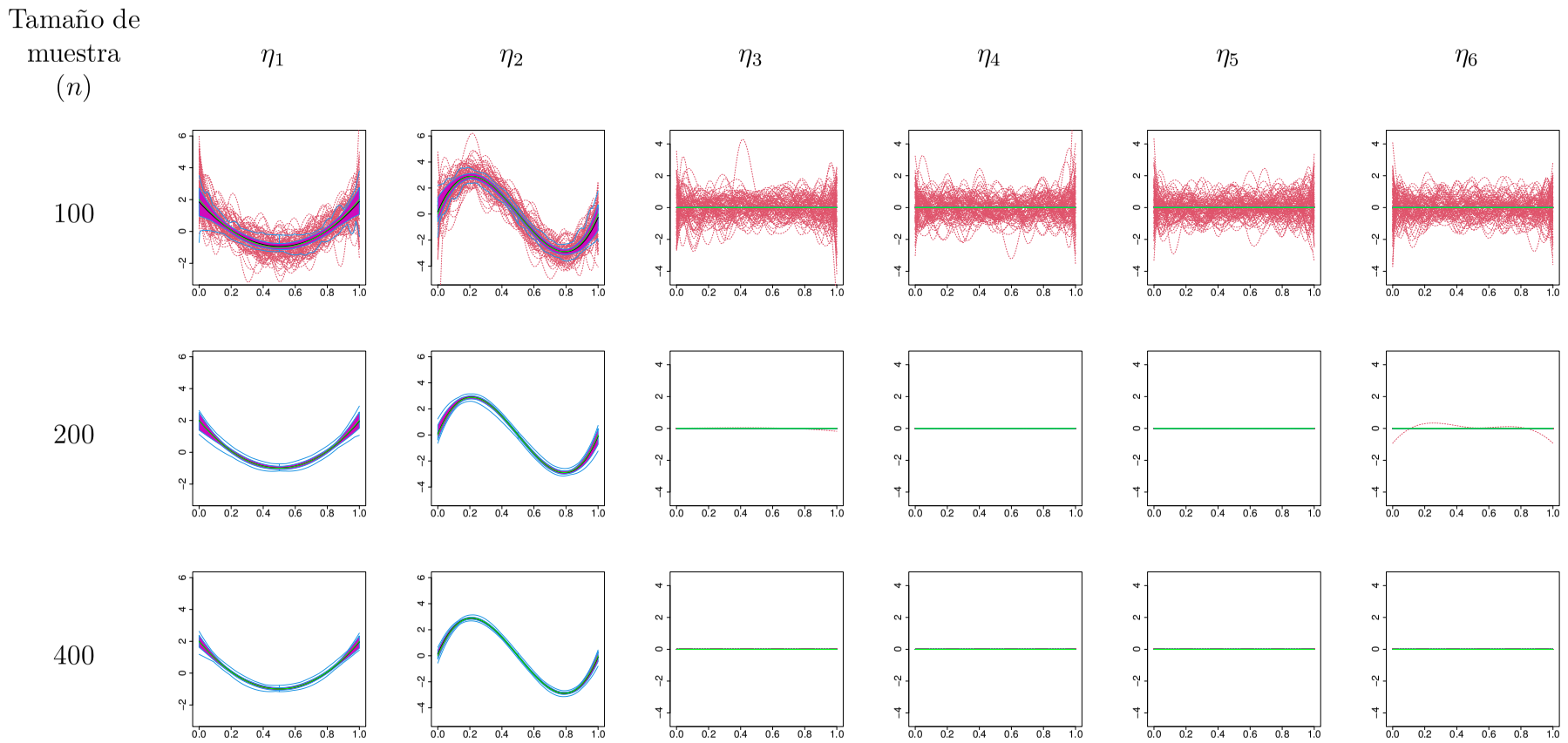

Figura 5.8: Boxplots funcionales para el Escenario III (η_1 a η_6).

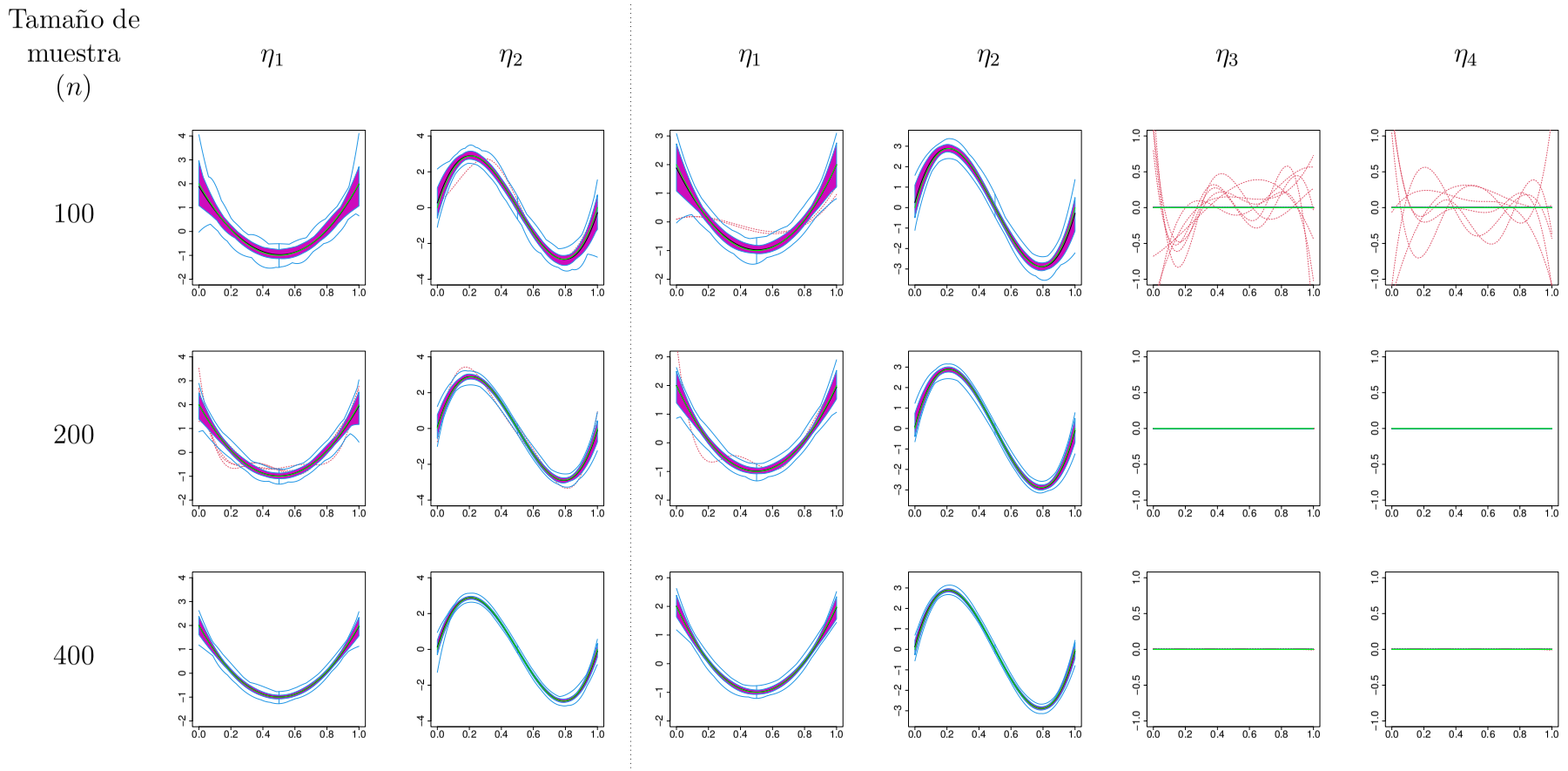

Figura 5.9: Boxplots funcionales: las dos primeras columnas corresponden al Escenario IV (η_1 a η_2) y las últimas cuatro columnas corresponden al Escenario V (η_1 a η_4).

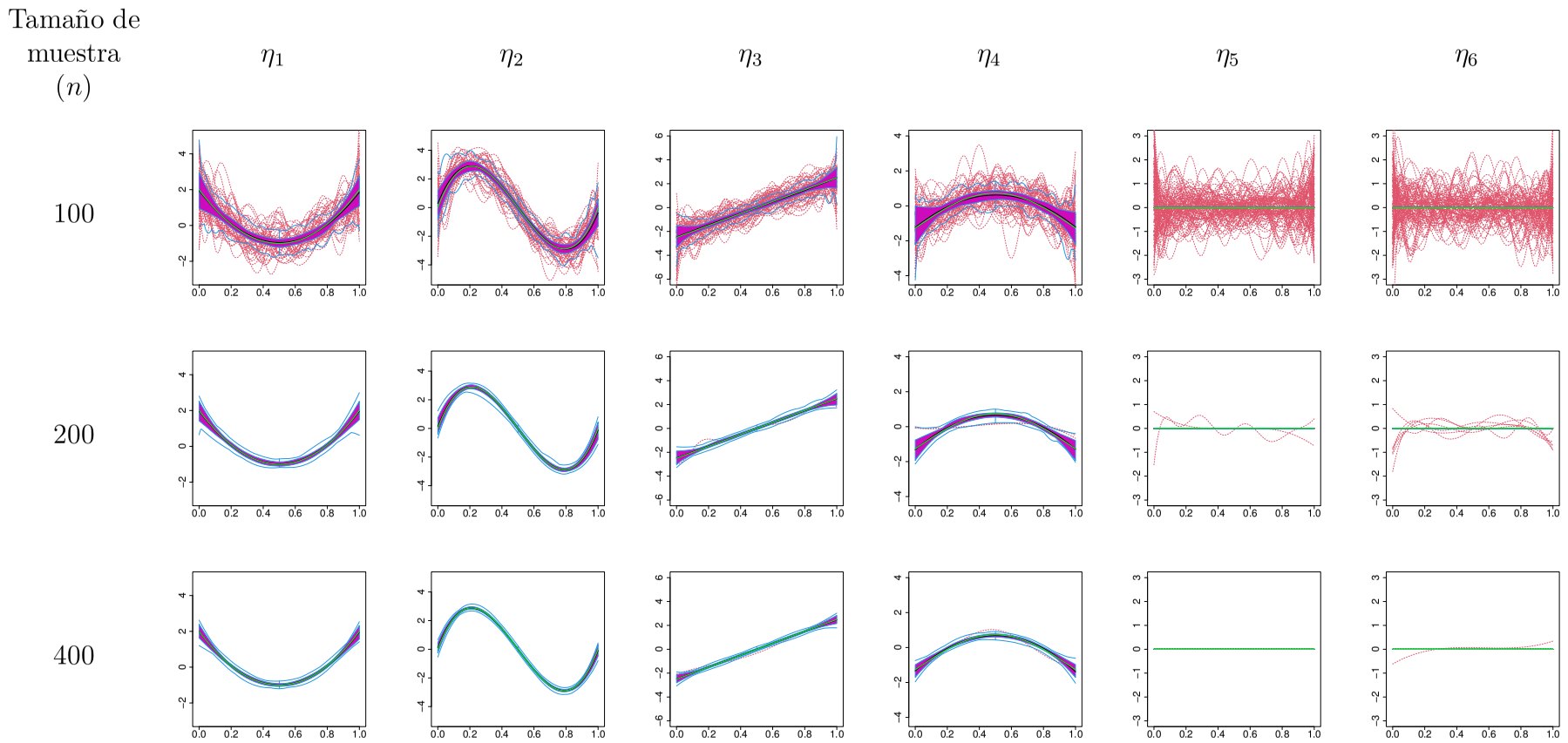

Figura 5.10: Boxplots funcionales para el Escenario VI (η_1 a η_6).

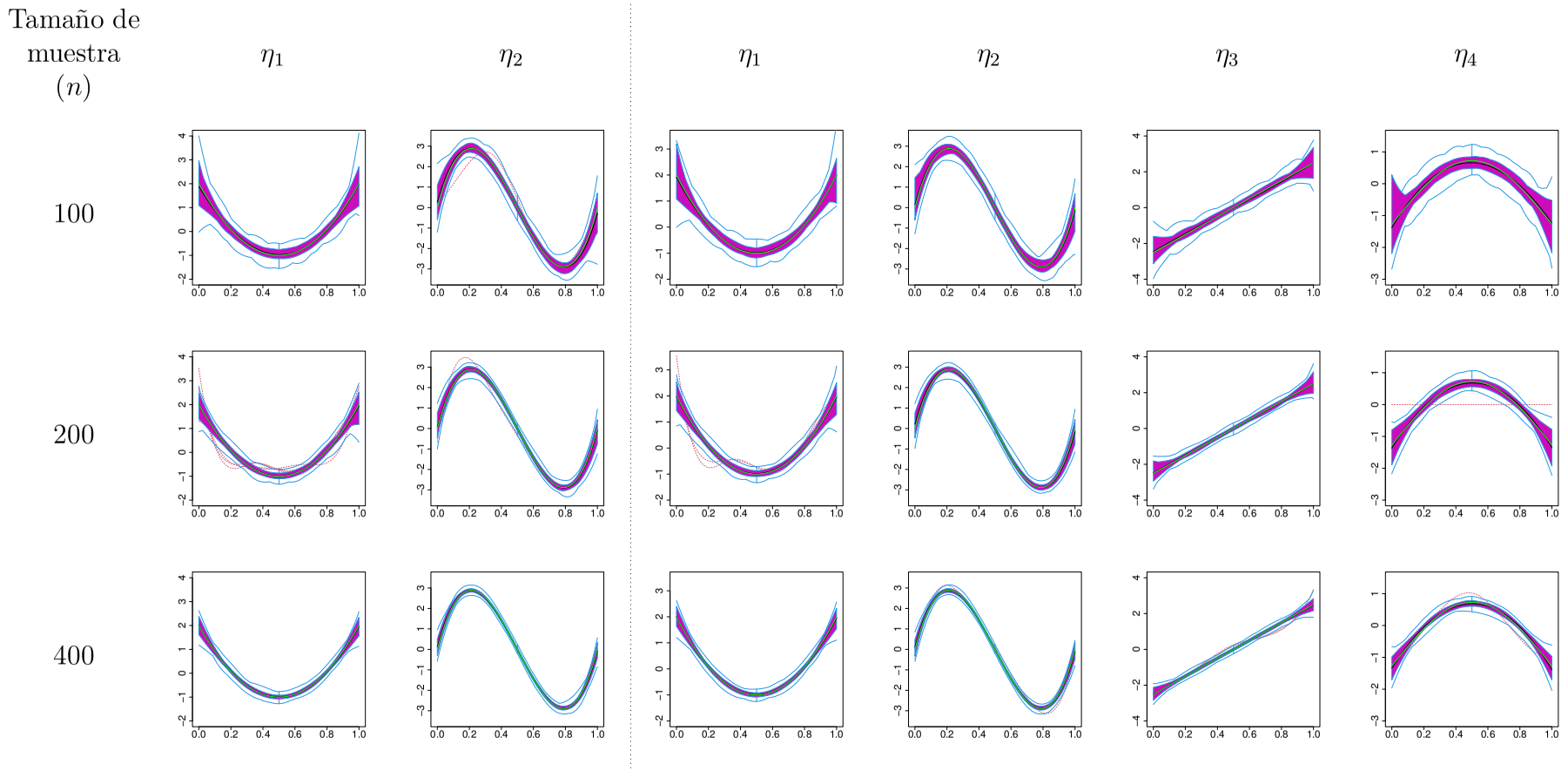

Figura 5.11: Boxplots funcionales: las dos primeras columnas corresponden el Escenario VII (η_1 a η_2) y las últimas cuatro columnas corresponden al Escenario VIII (η_1 a η_4).

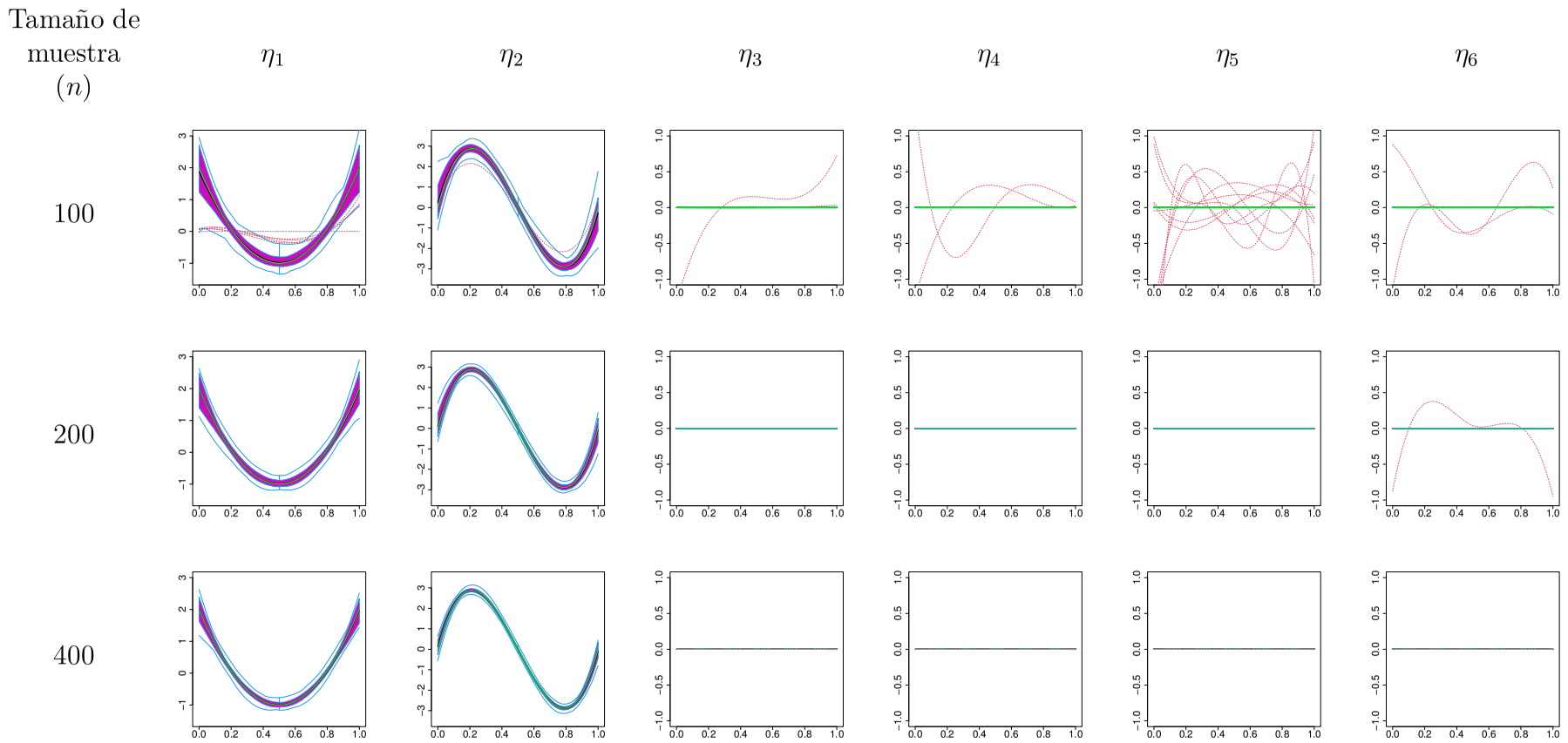

Figura 5.12: Boxplots funcionales para el Escenario IX (η_1 a η_6).

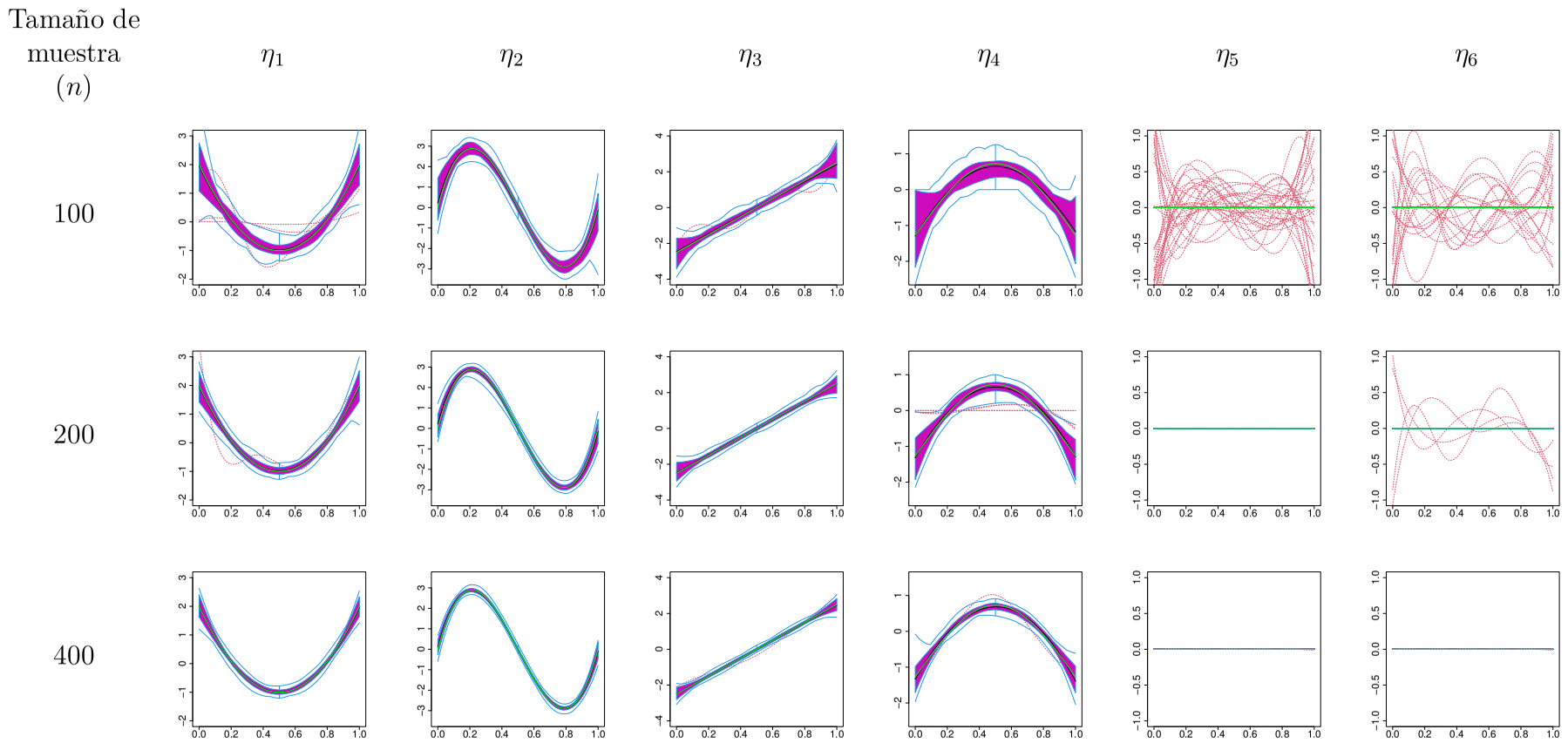

Figura 5.13: Boxplots funcionales para el Escenario X (η_1 a η_6).

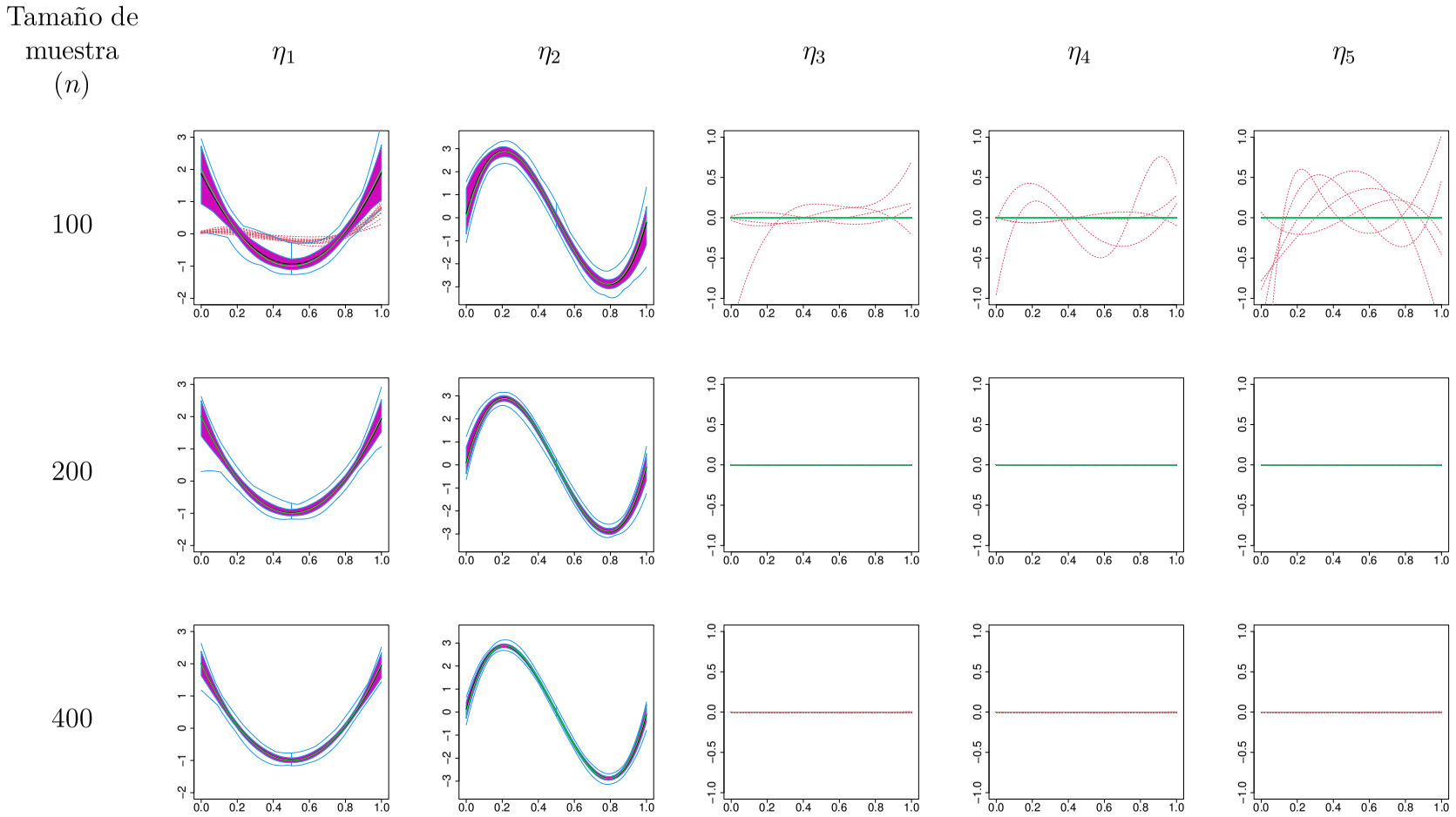

Figura 5.14: Boxplots funcionales para el Escenario XI (η_1 a η_5).

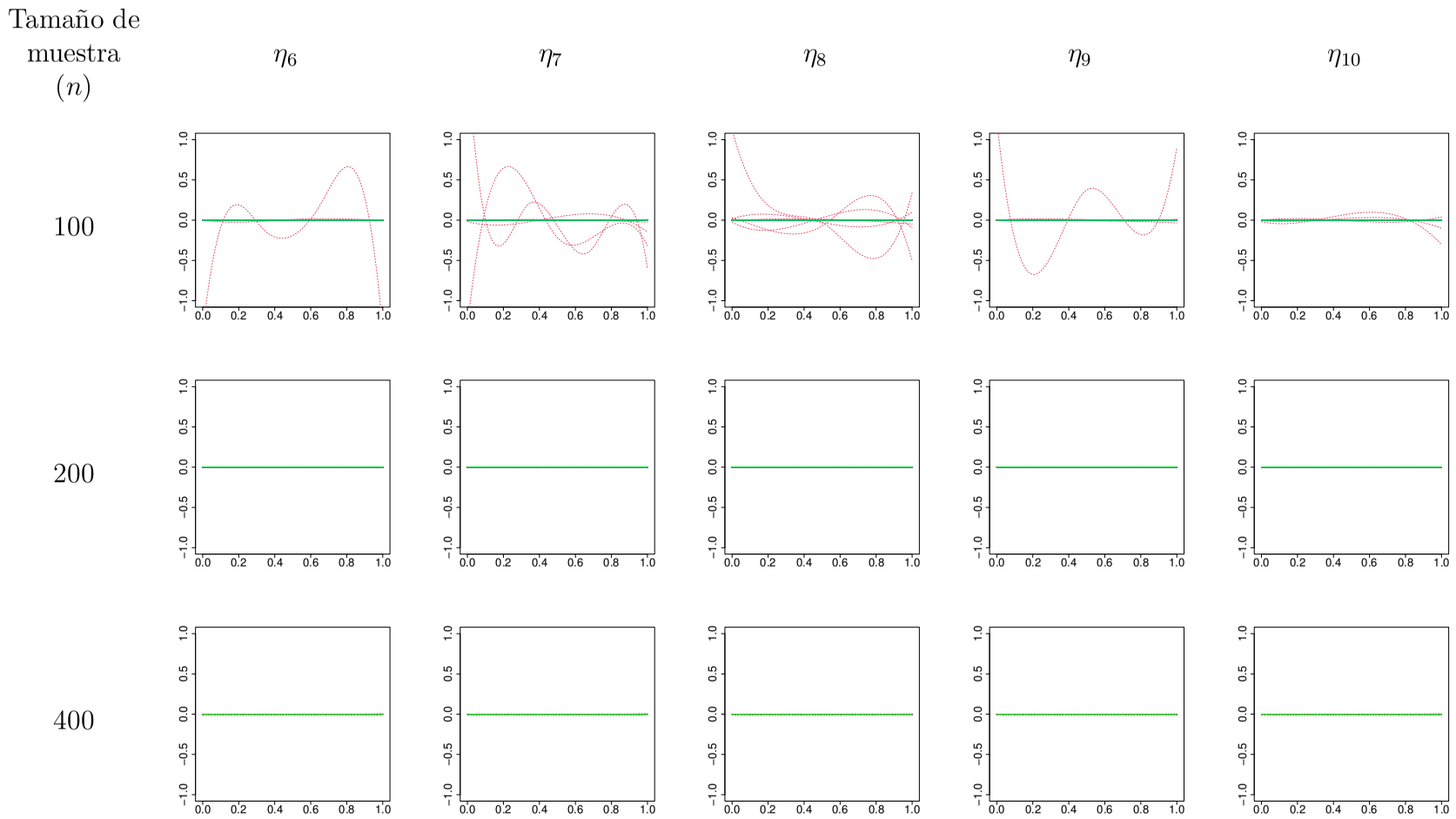

Figura 5.15: Boxplots funcionales para el Escenario XI (η_6 a η_{10}).

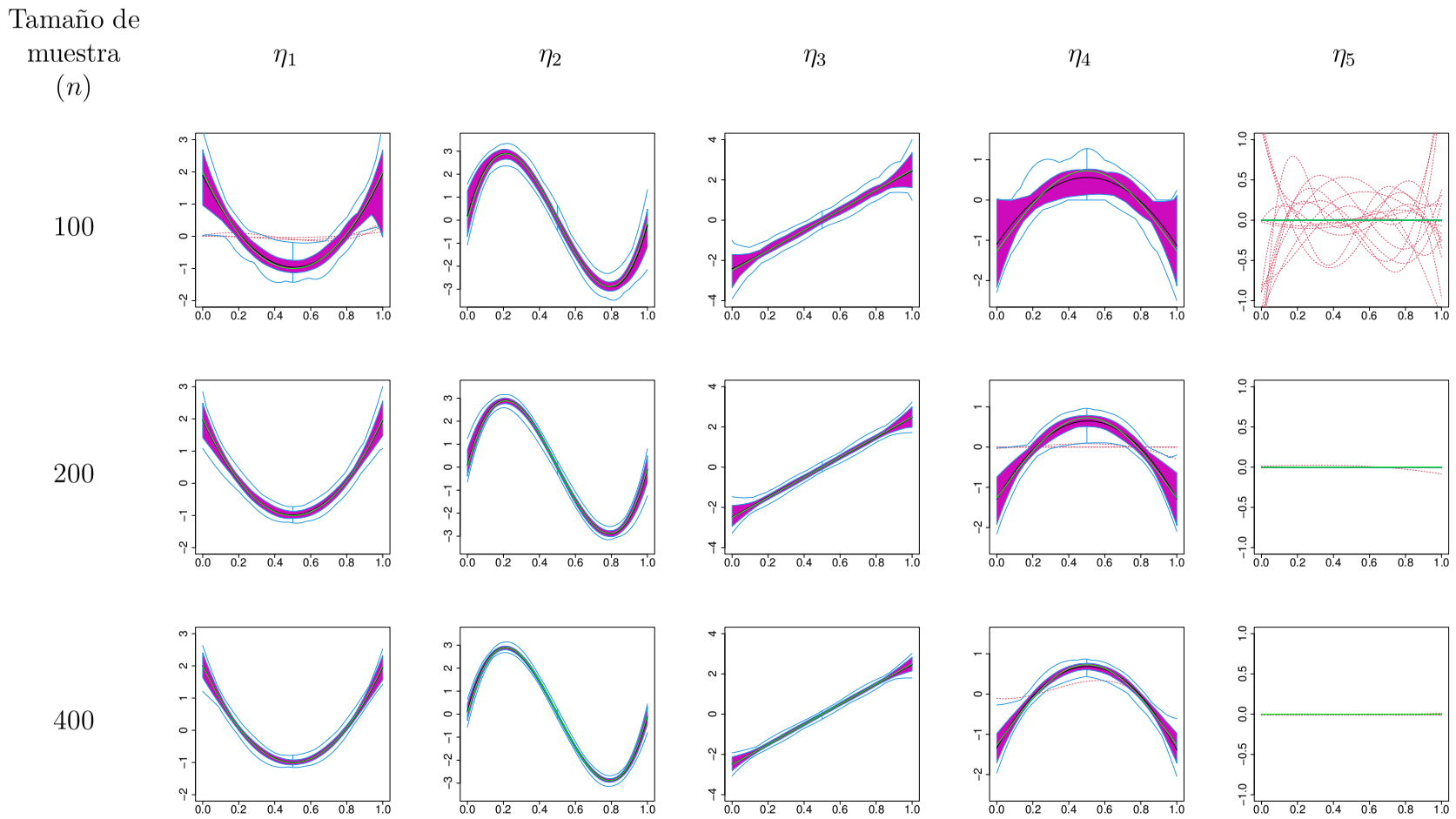

Figura 5.16: Boxplots funcionales para el Escenario XII (η_1 a η_5).

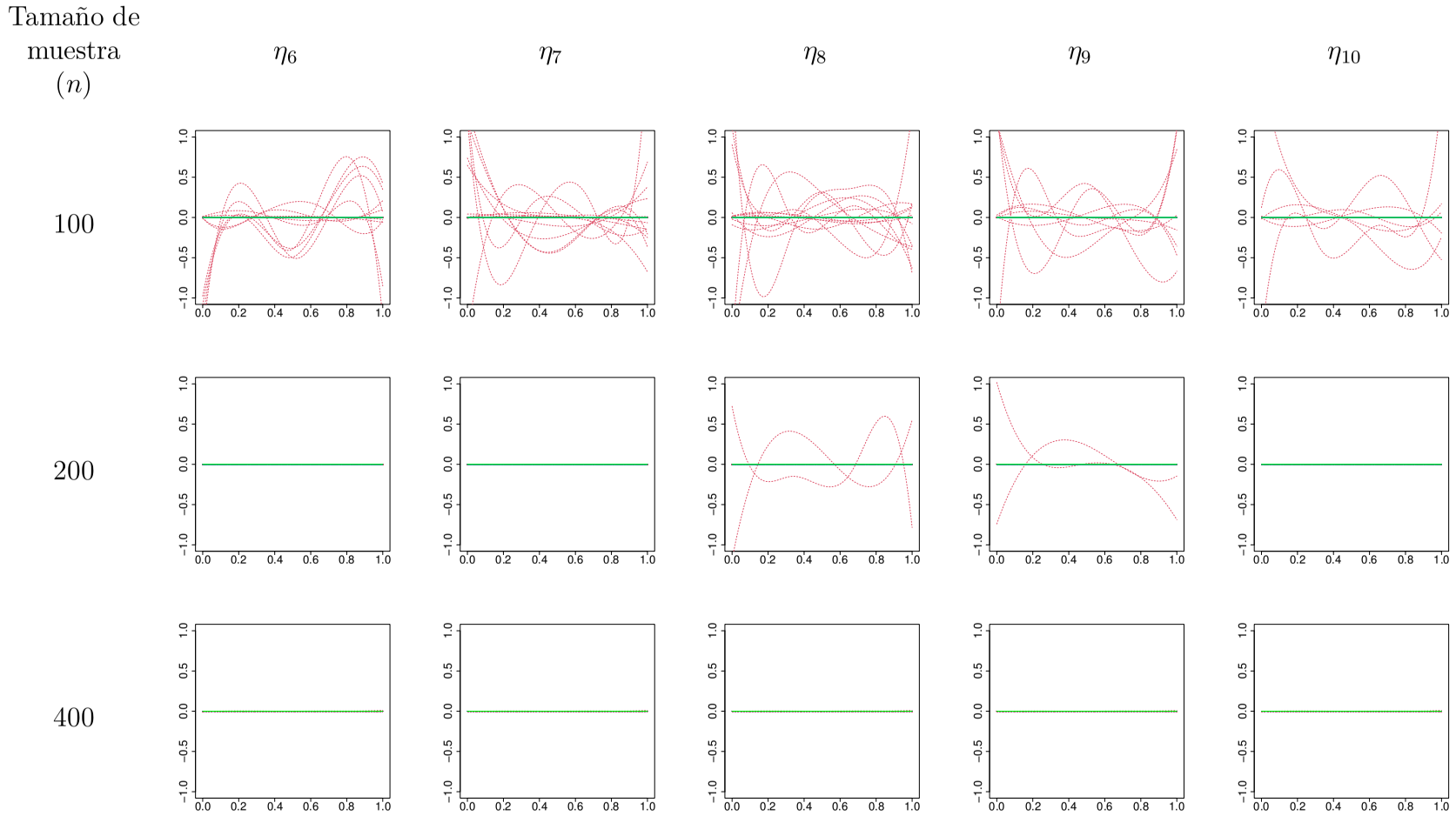

Figura 5.17: Boxplots funcionales para el Escenario XII (η_6 a η_{10}).

En términos generales, una primera observación es que, a medida que aumenta el tamaño muestral, se produce una notable contracción de las cajas y una reducción significativa en la longitud de los bigotes, lo cual indica una mejora en la precisión de las estimaciones. Además, las medianas funcionales se aproximan progresivamente a las funciones verdaderas, lo cual sugiere que el sesgo disminuye con el tamaño de muestra. Es importante destacar también que, en los casos donde la función verdadera es nula, si bien para $n = 100$ puede observarse una gran cantidad de curvas atípicas, esta cantidad se reduce drásticamente con $n = 200$ y prácticamente desaparece con $n = 400$, mostrando un comportamiento más estable y coherente con la capacidad del estimador para seleccionar variables.

Mirando las Figuras 5.7, 5.9 y 5.11 se observa que aquellos escenarios con pocas funciones activas ($p = 2$), como los Escenarios I, IV y VII, presentan en general un buen desempeño incluso con muestras pequeñas. En estos casos, las funciones estimadas ya se acercan bastante a las verdaderas con $n = 200$, y el ajuste se vuelve excelente con $n = 400$. Por otro lado, en las Figuras 5.8, 5.10, 5.12 y 5.14 a 5.17, asociadas a escenarios con mayor complejidad estructural, como los Escenarios III, VI, IX, XI y XII, que incluyen entre seis y diez funciones no nulas, imponen un desafío adicional. En $n = 100$ la estimación es más dispersa, los boxplots funcionales tienen cajas más anchas y hay mayor presencia de curvas atípicas, mostrando alguna pequeña dificultad en capturar el comportamiento de la función. Sin embargo, estos efectos se atenúan al incrementar el tamaño muestral. Es particularmente destacable que, incluso en los escenarios más exigentes, el procedimiento logra capturar adecuadamente tanto la forma general como el nivel de las funciones verdaderas cuando $n = 400$, con curvas medianas que se superponen visualmente con la función verdadera en verde. Este patrón se observa en funciones con diferentes niveles de suavidad, amplitud o curvatura, lo cual refuerza la capacidad del método para adaptarse a estructuras heterogéneas de las funciones aditivas.

Una mención especial merece la estimación de la función η_4, presente en los Escenarios VI, VIII, X y XII, ya que su forma trigonométrica, a diferencia de las funciones cuadrática, cúbica y lineal de η_1, η_2 y η_3, representa un desafío adicional para el procedimiento de estimación. En particular, en las Figuras 5.10, 5.11, 5.13 y 5.16, se observa que para $n = 100$ las cajas correspondientes a esta función son más amplias, y que la curva mediana (en negro) tiende a separarse ligeramente de la curva verdadera (en verde), lo cual sugiere una mayor dificultad relativa para capturar con precisión su forma oscilatoria. Aunque este efecto disminuye a medida que aumenta el tamaño muestral, se mantiene cierto rezago respecto al desempeño observado para las funciones polinómicas. Aun así, el estimador logra recuperar adecuadamente las principales características de η_4, lo cual habla de su capacidad de adaptación a distintas clases de funciones, aunque con un grado de precisión

algo menor en este caso particular.

En resumen, los resultados muestran que el estimador penalizado logra una gran precisión en la estimación de las componentes aditivas en todos los escenarios planteados, con un comportamiento especialmente satisfactorio al aumentar el tamaño muestral. La reducción simultánea del sesgo y de la variabilidad, la correcta identificación de funciones nulas, y la capacidad de ajuste en escenarios de alta dimensión, muestran evidencia del buen desempeño del método propuesto. Los patrones observados en los boxplots funcionales son coherentes con los resultados obtenidos en las medidas de error presentadas anteriormente, y aportan una clara visualización del poder de estimación del procedimiento, tanto en situaciones más sencillas como en escenarios más complejos.

5.4.3. Sobre la selección de los parámetros de penalización

En las Figuras 5.18, 5.19 y 5.20, se presentan los boxplots paralelos correspondientes a los parámetros de penalización auxiliares λ_1 y λ_2, según el criterio de BIC definido en la Sección 4.2, para cada uno de los escenarios, desglosados por tamaño de muestra. Cada figura corresponde a un grupo de cuatro escenarios de simulación: la Figura 5.18 contiene los Escenarios I a IV, la Figura 5.19 los Escenarios V a VIII y la Figura 5.20 los Escenarios IX a XII. En todas ellas, las filas representan los distintos escenarios y las columnas distinguen los parámetros λ_1 (a la izquierda) y λ_2 (a la derecha). Dentro de cada celda, se incluyen diagramas de caja paralelos que resumen los valores obtenidos para $n = 100$, $n = 200$ y $n = 400$, separados visualmente mediante líneas punteadas. Las cajas correspondientes a λ_1 se colorean en azul, mientras que las asociadas a λ_2 se muestran en verde. Esta organización permite visualizar de forma clara cómo evolucionan los parámetros seleccionados de la parte lineal y de la parte aditiva al variar, tanto el escenario como el tamaño de muestra, facilitando la comparación entre los distintos casos. Recordemos que la selección de los parámetros se realiza sobre una grilla $\mathcal{G} = \{(\ell_1, \ell_2) : \ell_1 \in \Lambda_1, \ell_2 \in \Lambda_2\}$, donde $\Lambda_1 = \Lambda_2$ contiene valores entre 0 y 2 distribuidos de forma no uniforme, con mayor densidad en la vecindad del cero. Esta distribución busca capturar adecuadamente valores pequeños, relevantes cuando se requiere escasa penalización, sin descuidar la posibilidad de seleccionar valores moderados o grandes cuando el ajuste lo exige.

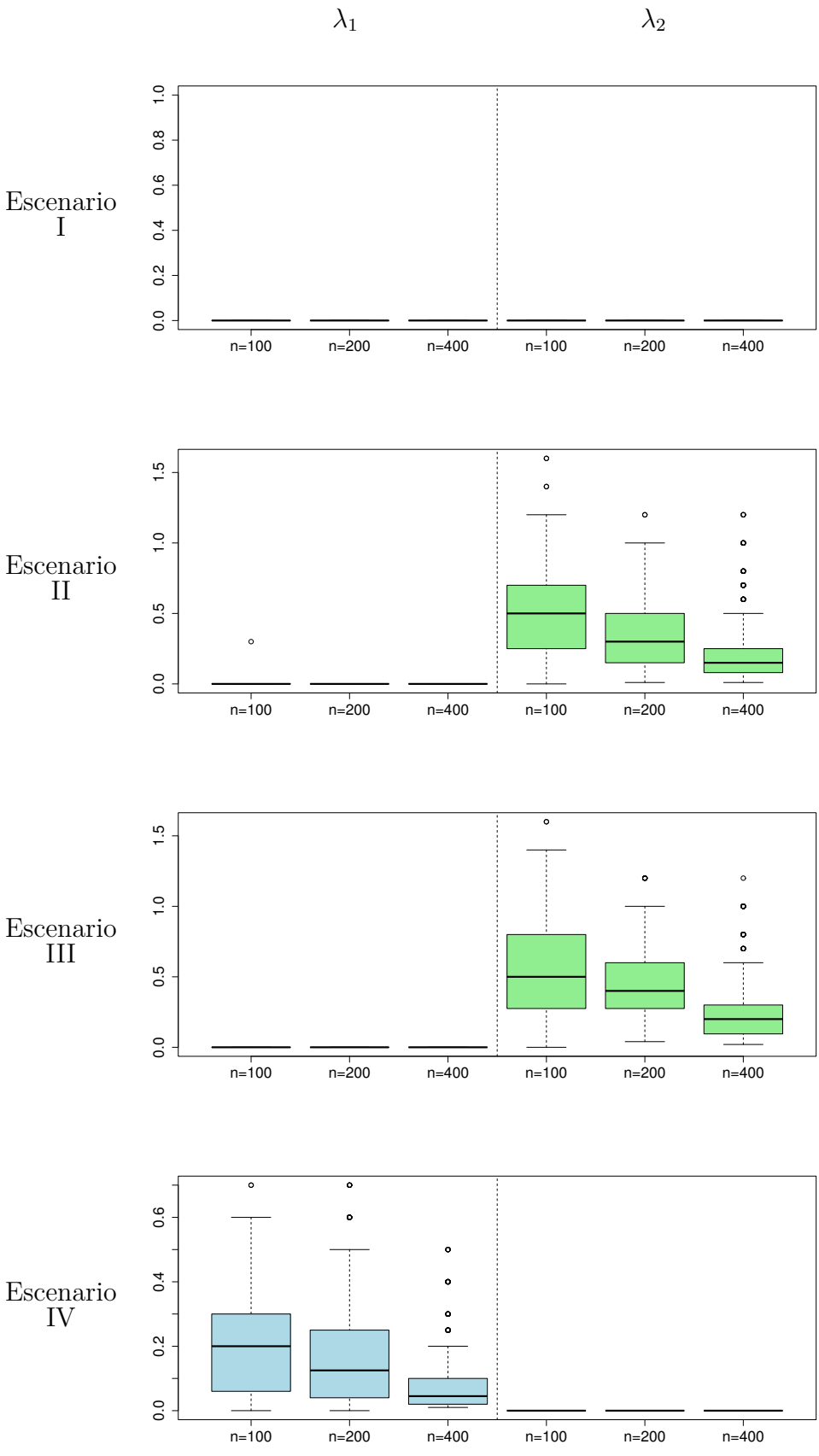

Figura 5.18: Boxplots paralelos de los valores de λ_1 y λ_2 seleccionados para los tres tamaños de muestra considerados, en los Escenarios I a IV.

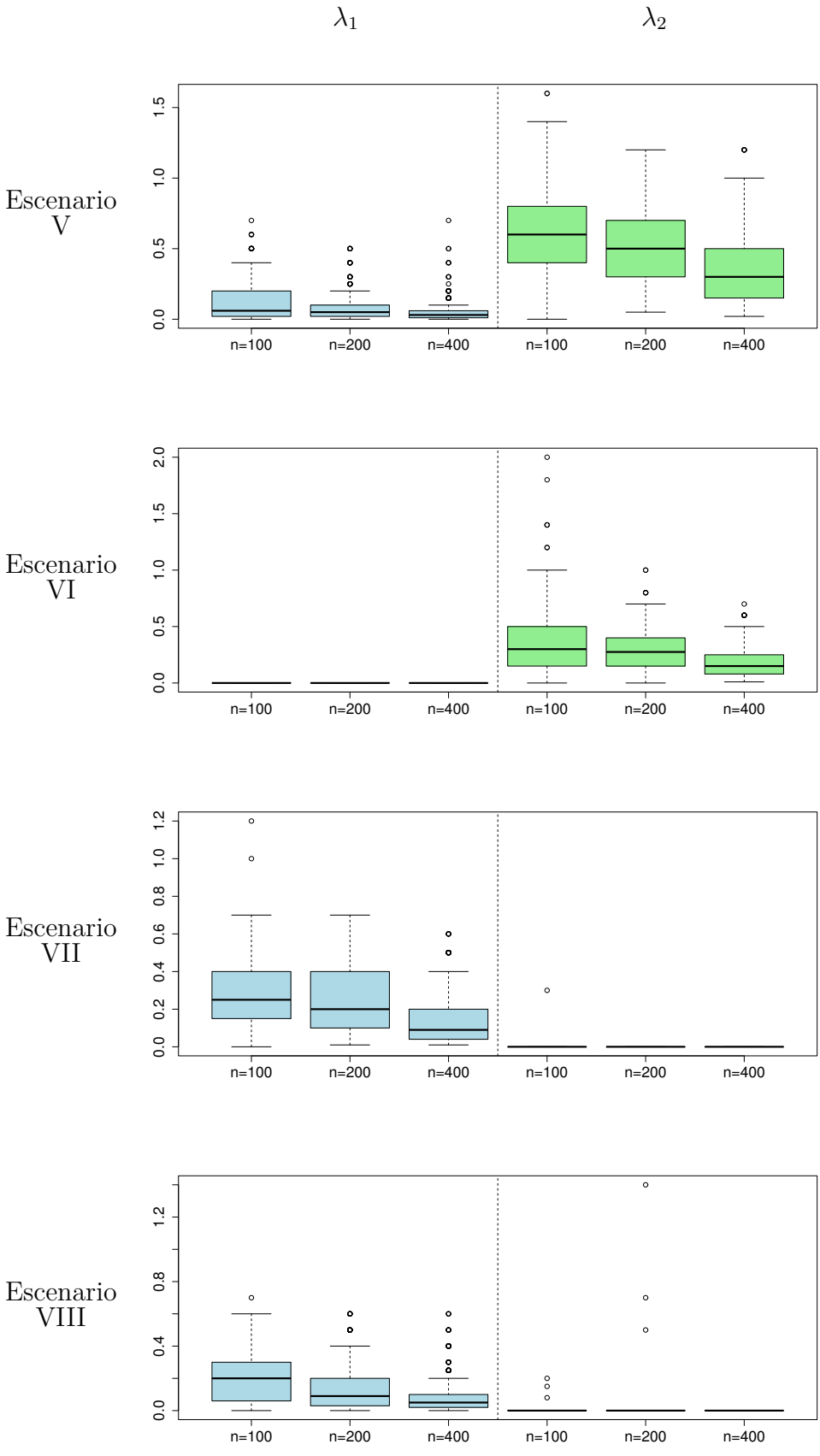

Figura 5.19: Boxplots paralelos de los valores de λ_1 y λ_2 seleccionados para los tres tamaños de muestra considerados, en los Escenarios V a VIII.

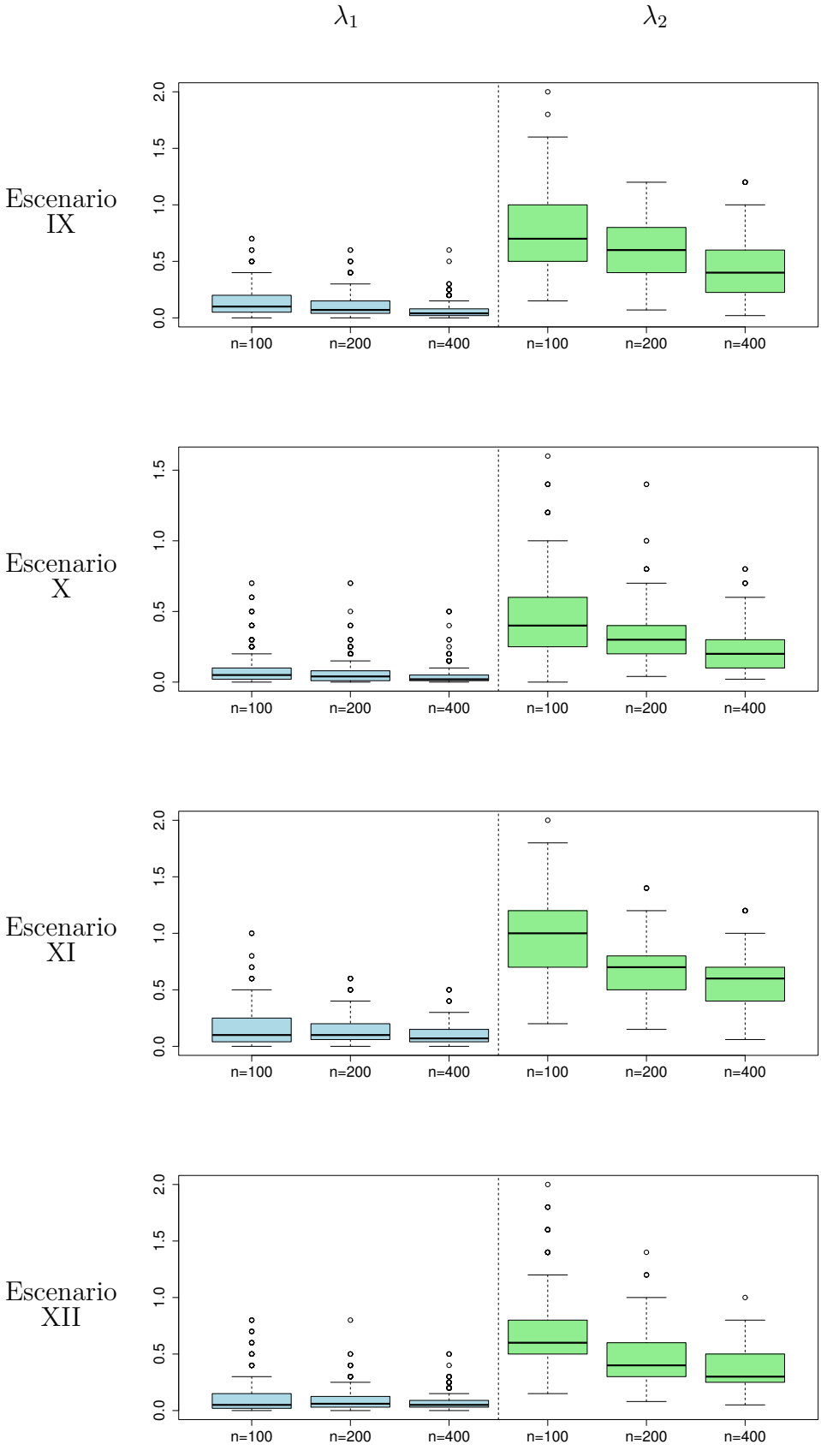

Figura 5.20: Boxplots paralelos de los valores de λ_1 y λ_2 seleccionados para los tres tamaños de muestra considerados, en los Escenarios IX a XII.

Como puede observarse en las Figuras 5.18 y 5.19, los valores seleccionados de λ_1 tienden a concentrarse cerca de cero en varios escenarios. Esta situación es especialmente marcada en los Escenarios I, II, III y VI, donde todas las variables de la parte lineal son relevantes. En dichos casos, la gran mayoría de las muestras generadas conducen a valores nulos o cercanos a cero, lo que provoca que los boxplots colapsen en un único valor y no presenten la forma habitual de caja. En general, para estos primeros ocho escenarios, incluso cuando aparecen valores distintos de cero, éstos son moderados: registrándose el más alto como un outlier en el Escenario IV para $n = 100$ un valor de λ_1 de 0.7. En los Escenarios V, VII y VIII de la Figura 5.19 y los escenarios IX a XII de la Figura 5.20, los boxplots de λ_1 muestran algo más de variación, con cajas visibles ubicadas en la parte inferior del gráfico, vallas a derecha y a la izquierda y algunos outliers a derecha. Los valores de λ_1 registrados disminuyen con el tamaño de muestra, pudiéndose observar esto en cajas cada vez más pequeñas y concentradas más cerca del 0 a medida que n crece.

En cuanto a λ_2, los resultados muestran una mayor variabilidad en comparación con λ_1, seleccionando valores más grandes de este parámetro que para el caso lineal. Sin embargo, el análisis es parecido. En primer lugar y tal como puede observarse en las Figuras 5.18 y 5.19, en los Escenarios I, IV, VII y VIII, donde todas las funciones aditivas son relevantes, los boxplots colapsan en cero, tal como ocurría cuando todas las componentes de la parte lineal eran significativas. No obstante, en algunos casos se observan valores atípicos alejados de cero: por ejemplo, en el Escenario VIII con $n = 200$, se registra una replicación para la que fue seleccionado $\lambda_2 = 1.4$. En los restantes escenarios en los que el estimador sí debe seleccionar variables, los boxplots obtenidos presenten la forma tradicional del boxplot, pudiéndose observar, nuevamente y tal como ocurría para el caso del parámetro λ_1, cajas ubicadas en la parte inferior del gráfico, vallas de ambos lados, asimetría a derecha y la presencia de algunos datos atípicos. Al igual que antes, a medida que el tamaño de muestra aumenta, el procedimiento selecciona valores del parámetro más pequeño, pudiéndose notar esto en cajas más pequeñas y concentradas cerca del 0.

En lo que respecta a la relación entre λ_1 y λ_2, es posible observar que, sistemáticamente, cuando el procedimiento sí tiene que seleccionar variables, los valores de λ_1 seleccionados son más pequeños que los valores seleccionados de λ_2. Esto puede verse, por ejemplo, en los Escenarios IX, XI y XII, en los que los boxplots de λ_2 obtenidos se encuentran por arriba de las cajas de los boxplots obtenidos para λ_1. Esto sugiere que el procedimiento propuesto requiere, en líneas generales, de valores del parámetro λ_2 más grandes que los requeridos para λ_1 para alcanzar un buen compromiso entre estimación y selección de variables. Es interesante observar que los valores seleccionados para el parámetro λ_1 por el criterio de BIC propuesto,

para los tres valores de muestra considerados, son, en general, menores a 0.8. Esto respalda la decisión de utilizar una grilla más fina cerca del cero (y hasta el 0.80) y expandida de manera creciente, permitiendo al procedimiento explorar de forma precisa un rango relevante de posibles valores tanto para λ_1 como para λ_2.

5.4.4. Sobre la selección de los parámetros de suavizado

En las Figuras 5.21, 5.22 y 5.23 se presentan histogramas que muestran la frecuencia con la que se seleccionó cada valor del parámetro de suavizado k a lo largo de las 500 replicaciones realizadas. Recordemos este parámetro determina la cantidad de funciones base utilizadas en la representación spline de cada componente aditiva. Cada figura está organizada de forma que las filas corresponden a los diferentes escenarios considerados, mientras que las columnas separan los tamaños muestrales $n = 100$, $n = 200$ y $n = 400$. La Figura 5.21 contiene los resultados correspondientes a los Escenarios I a IV, la Figura 5.22 muestra los Escenarios V a VIII, y la Figura 5.23 presenta los resultados para los Escenarios IX a XII. Esta disposición permite evaluar cómo varía la selección de k según la complejidad del escenario y el tamaño de muestra, así como detectar posibles patrones de estabilización o dispersión en la elección del grado de suavizado.

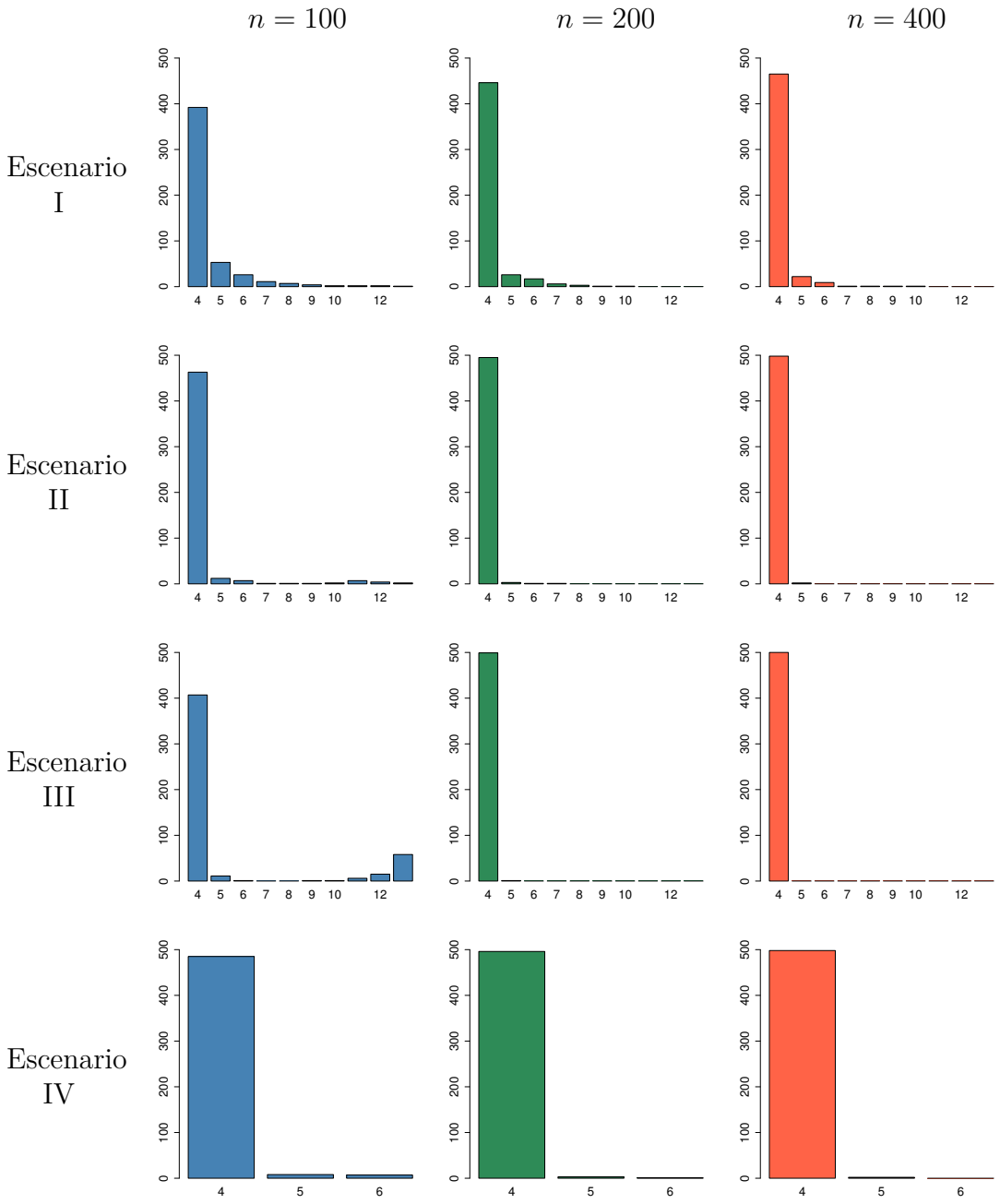

Figura 5.21: Gráficos de barras de valores de k para los tres tamaños de muestra considerados, en los Escenarios I a IV.

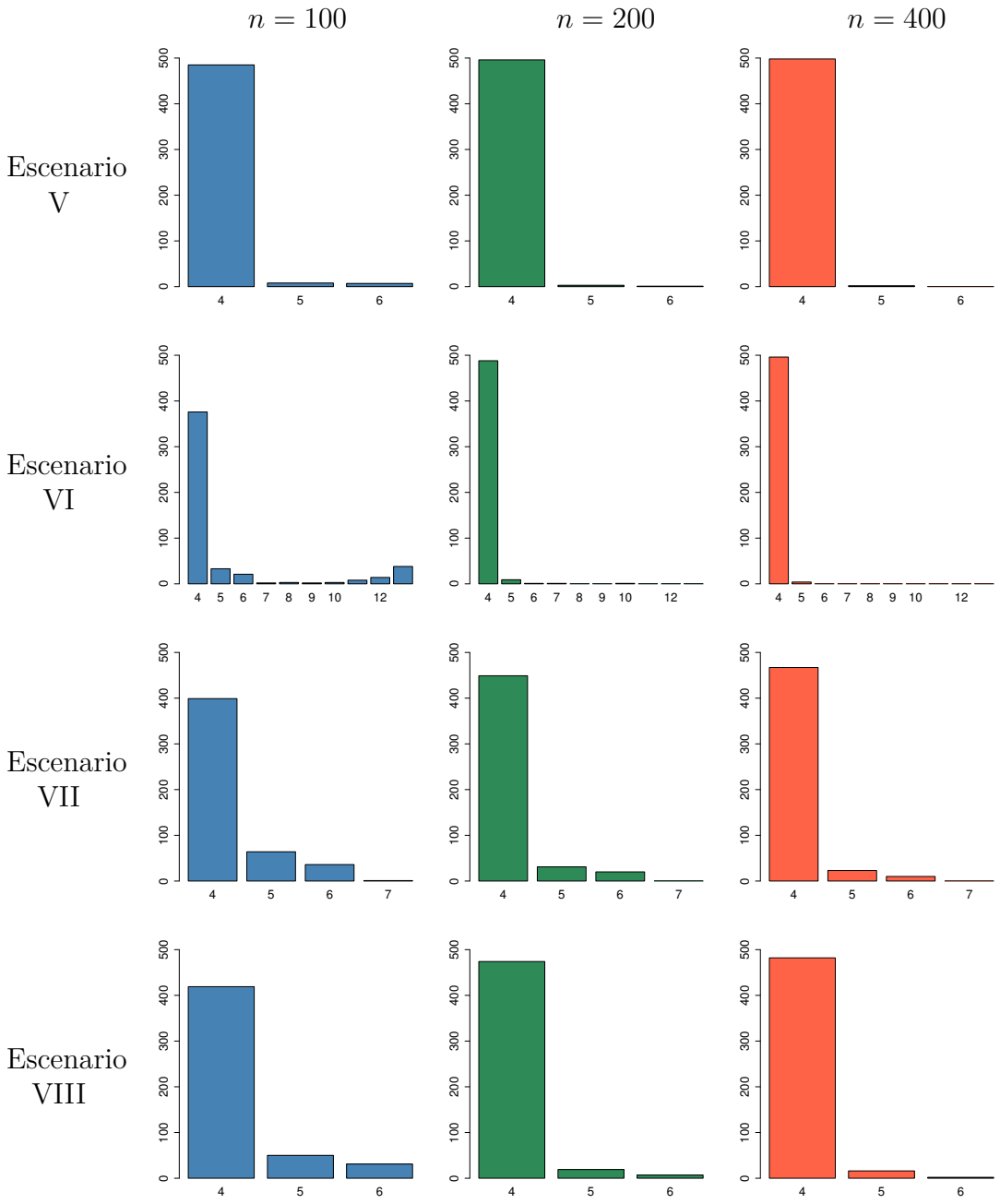

Figura 5.22: Gráficos de barras de valores de k para los tres tamaños de muestra considerados, en los Escenarios V a VIII.

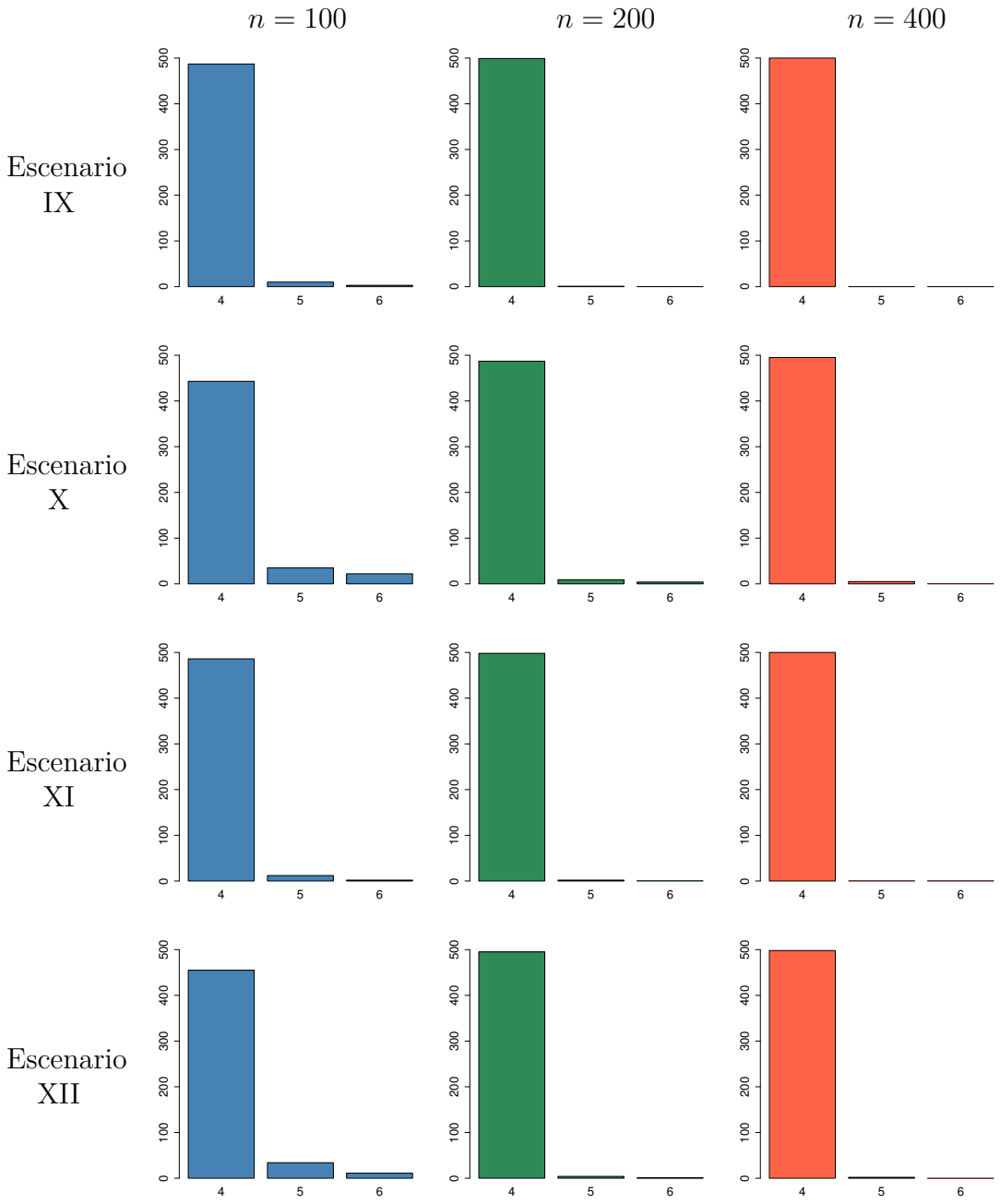

Figura 5.23: Gráficos de barras de valores de k para los tres tamaños de muestra considerados, en los Escenarios IX a XII.

El análisis de las Figuras 5.21, 5.22 y 5.23 revela un patrón consistente en la selección del parámetro de suavizado k: a medida que el tamaño muestral aumenta, la variabilidad en los valores seleccionados disminuye notablemente. Este comportamiento es coherente con lo observado anteriormente para los parámetros de penalización λ_1 y λ_2, y sugiere que el procedimiento propuesto se estabiliza a medida que n aumenta. Por ejemplo, en la Figura 5.21, correspondiente a los Escenarios I a IV, puede observarse que en el Escenario I, para $n = 100$, si bien el valor $k = 4$ es el más frecuentemente seleccionado, el procedimiento llegó a elegir todo el rango permitido, alcanzando incluso valores extremos como $k = 13$. En cambio, para $n = 200$ y especialmente para $n = 400$, la selección de k se concentra en valores más bajos, restringiéndose aproximadamente al rango 4 a 10, y aumentando significativamente la frecuencia de selección de $k = 4$. Este patrón de concentración se repite en los demás escenarios de la figura, indicando que el procedimiento tiende a elegir valores más bajos de k a medida que aumenta el tamaño de muestra. La Figura 5.22, que presenta los Escenarios V a VIII, confirma esta tendencia. Para $n = 100$, sólo en el Escenario VI se observa un rango amplio de valores de k, mientras que en los Escenarios V, VII y VIII la selección ya se restringe a rangos más reducidos. En particular, en los Escenarios VII y VIII, los valores seleccionados se limitan a $k = 4, 5, 6$ y, en el caso del Escenario VIII, el valor $k = 4$ es el más elegido incluso en muestras pequeñas. Esta concentración se acentúa para $n = 200$ y se vuelve muy marcada en $n = 400$, donde la mayoría de las réplicas coinciden en seleccionar $k = 4$. Por último, en la Figura 5.23, correspondiente a los Escenarios IX a XII, este patrón es aún más pronunciado. Incluso para $n = 100$, la selección de k se concentra en el rango 4 a 6, sin alcanzar valores extremos. A medida que aumenta el tamaño muestral, la selección converge sistemáticamente hacia 4, que se vuelve el valor predominante en todos los escenarios de esta figura.

Capítulo 6

Análisis de datos reales

Los estudios observacionales han sugerido que concentraciones plasmáticas reducidas de betacaroteno podrían estar asociadas con un mayor riesgo de desarrollar ciertos tipos de cáncer. Por ello, se diseñó un estudio para investigar la relación entre las concentraciones plasmáticas de betacaroteno y diversos factores regulatorios, tales como características personales (sexo y edad) y ciertos factores dietéticos (consumo de alcohol, tabaquismo, entre otros). El conjunto de datos `plasma`, recolectado por Nierenberg et al. (1989) y disponible en el paquete `gamlss.data` de `R`, contiene 315 observaciones y 14 variables. Es de interés investigar la relación entre la variable respuesta, denotada BETACAROTENO, y las covariables que describiremos a continuación:

- AGE: edad en años.
- SEX: sexo (1=hombre; 2=mujer).
- SMOKSTAT: estado de fumador (1=nunca; 2=ex fumador; 3=actualmente fumador).
- BMI: índice de masa corporal (peso/(altura)^2).
- VITUSE: uso de Vitamina (1=sí, bastante a menudo; 2= sí, no tan frecuente; 3=no).
- CALORIES: número de calorías consumidas por día.
- FAT: gramos de grasa consumidos por día.
- FIBER: gramos de fibra consumidos por día.

- ALCOHOL: número de tragos con alcohol consumidos por semana.
- CHOL: colesterol consumido (mg/día).
- BETADIET: betacaroteno dietético consumido (mcg/día).

Tal como remarcaron Liu et al. (2011) y Guo et al. (2013), hay un punto de alto leverage en el consumo de alcohol que, al igual que estos autores, será excluído del análisis. Además, las variables categóricas fueron convertidas en variables *dummy* (o variables indicadoras) previo ingresar al modelo. A saber, SMOKSTAT entra en el modelo como SMOK1 (1=ex fumador; 0=otro) y SMOK2 (1=actualmente fumador; 0=other), mientras que VITUSE se separa en VIT1 (1=sí, bastante a menudo; 0=otro) y VIT2 (1=sí, no tan frecuente; 0=otro). Todas las variables, a excepción de las dummies, fueron estandarizadas restándoseles la media y dividiendo por su desvío estándar muestral.

Este conjunto de datos ha sido ampliamente estudiado por varios autores, especialmente en el contexto de modelos semiparamétricos y, en particular, de modelos parcialmente lineales aditivos. Por ejemplo, en Liu et al. (2011) se utilizó un enfoque basado en mínimos cuadrados para modelar el logaritmo de BETAPLASMA con nueve covariables: BMI, CALORIES, FAT, FIBER, BETADIET, SEX, ALCOHOL, SMOK1 y SMOK2 en la parte lineal; y AGE y CHOL en la parte aditiva. El mismo modelo fue considerado en Banerjee y Ghosal (2014), pero utilizando un método bayesiano para la selección de variables. Por su parte, Jeong et al. (2022) también aplicaron un enfoque bayesiano, pero enfocado en estudiar qué variables podrían incluirse en la componente lineal y cuáles en la aditiva. Por otra parte, Guo et al. (2013) propusieron un enfoque de regresión por cuantiles compuestos con un procedimiento de regularización para seleccionar variables de la componente lineal, modelando BETAPLASMA previamente estandarizada. Incluyeron diez covariables en la parte lineal: SEX, SMOK1, SMOK2, BMI, VIT1, VIT2, CALORIES, FAT, ALCOHOL y BETADIET, mientras que AGE, CHOL y FIBER fueron incluidas en la parte aditiva. Todas las covariables, tanto de la parte lineal como de la parte aditiva, fueron estandarizadas restándoles a cada una la media muestral y dividiendo por el desvío estándar muestral. Por otro lado, Lv et al. (2017) consideraron un enfoque de regresión modal para modelar BETAPLASMA con diez covariables en la parte lineal: SEX, SMOKSTAT, QUET, VITUSE, CALORIES, FAT, ALCOHOL, BETADIET y dos covariables adicionales llamadas RETDIET y RETPLASMA, manteniendo las mismas tres covariables (AGE, FIBER y CHOL) en la parte aditiva del modelo. Aunque en su análisis trataron a SMOKSTAT como una variable discreta, convertirla en variables categóricas podría haber estado más en línea con las prácticas estándar de modelado. Finalmente, Boente y Martínez (2024) estudiaron el mismo conjunto de datos y propusieron el mismo modelo

que Guo et al. (2013), pero aplicando la selección de variables únicamente a la componente lineal del modelo.

De esta forma, consideramos el mismo modelo que en Guo et al. (2013) y en Boente y Martínez (2024) en el que se modela los niveles plasmáticos de betacaroteno con las covariables AGE, CHOL y FIBER entrando al modelo de forma aditiva, mientras que las demás variables entran de forma lineal. Más precisamente, el modelo parcialmente lineal aditivo considerado es el siguiente:

$$\begin{aligned}\text{BETAPLASMA} = \mu &+ \beta_1\text{SEX} + \beta_2\text{SMOK1} + \beta_3\text{SMOK2} + \beta_4\text{BMI} + \beta_5\text{VIT1} \\ &+ \beta_6\text{VIT2} + \beta_7\text{CALORIES} + \beta_8\text{FAT} + \beta_9\text{ALCOHOL} \\ &+ \beta_{10}\text{BETADIET} + \eta_1(\text{AGE}) + \eta_2(\text{CHOL}) + \eta_3(\text{FIBER}) + u,\end{aligned}$$

donde los errores u se suponen independientes de las covariables. De esta manera, y según la notación utilizada en este trabajo, resulta que las variables que ingresan al modelo de forma lineal son $\mathbf{Z} = (\text{SEX}, \text{SMOK1}, \text{SMOK2}, \text{BMI}, \text{VIT1}, \text{VIT2}, \text{CALORIES}, \text{FAT}, \text{ALCOHOL}, \text{BETADIET})^{\mathrm{T}} \in \mathbb{R}^{10}$ y que las variables aditivas son $\mathbf{X} = (X_1, X_2, X_3)^{\mathrm{T}} = (\text{AGE}, \text{CHOL}, \text{FIBER})^{\mathrm{T}} \in \mathbb{R}^3$.

Para el ajuste del modelo, estimamos las componentes aditivas usando splines cúbicos. En la estimación de η_j, para $1 \leq j \leq 3$, los nodos internos son tomados como los cuantiles $\ell/(N_j + 1)100\,\%$ de los valores observados de las covariables X_j, para $\ell = 1, \ldots, N_j$. El número de nodos internos (N_j) fue considerado el mismo para cada una de las tres covariables que ingresan la parte aditiva, es decir, se consideró $k_j = k$ para $j = 1, 2, 3$.

Los parámetros de penalización auxiliares λ_1 y λ_2 y el parámetro de suavizado k son seleccionados automáticamente usando los criterios de BIC desarrollados en este trabajo (ver Secciones 4.2 y 4.3, respectivamente). El parámetro de suavizado k fue seleccionado en la grilla $\{4, \ldots, 13\}$. En cuanto a la grilla para los parámetros de penalización, para $\widetilde{\boldsymbol{\lambda}} = (\lambda_1, \lambda_2)$, consideramos la grilla $\mathcal{G} = \{(\ell_1, \ell_2) : \ell_1 \in \Lambda_1, \ell_2 \in \Lambda_2\}$, donde $\Lambda_1 = \{0, 0.01, 0.02, \ldots, 0.20, 0.25, \ldots, 2\}$ y $\Lambda_2 = \{0, 0.05, 0.10, \ldots, 1, 1.1, 1.2, \ldots, 2, 2.4, 2.8, \ldots, 10, 15, 20, \ldots, 250\}$. Mientras que la grilla Λ_1 para el parámetro λ_1 va desde 0 hasta 0.20 con paso 0.01 y desde 0.20 hasta 2 con paso 0.05, la grilla Λ_2 para el parámetro λ_2 consiste en una primera parte más fina, que tiene paso 0.05 hasta el 1, luego tiene paso 0.1 hasta el 2, sigue con paso 0.4 hasta el 10 y desde ahí con paso 5 hasta el 250.

Para medir la calidad de predicción del estimador, 100 observaciones fueron separadas aleatoriamente para constituir el conjunto de testeo, mientras que las restantes 214 fueron asignadas al conjunto de entrenamiento con el objetivo de estimar y seleccionar variables. Es decir, separamos la muestra en una relación aproximada de $70-30$, siendo un $70\,\%$ de los datos considerados como muestra de entrenamien-

to y un 30 % como muestra de testeo. El rendimiento de la predicción fue medido a través del Error de Predicción Absoluto Mediano, (MAPE, *Median Absolute Prediction Error*), definido como la mediana de $\{|y_i - \widehat{y}_i|, \quad i = 1, \ldots, 100\}$, donde $(y_i, \mathbf{z}_i^{\mathrm{T}}, \mathbf{x}_i^{\mathrm{T}})^{\mathrm{T}}, 1 \leq i \leq 100$ son las observaciones del conjunto de testeo y $\widehat{y}_i$ es la predicción obtenida por el estimador basado en las 214 observaciones restantes en $(\mathbf{z}_i^{\mathrm{T}}, \mathbf{x}_i^{\mathrm{T}})^{\mathrm{T}}$. Este procedimiento se replicó $N_{rep} = 50$ veces. Finalmente, se calculó el promedio de los 50 valores obtenidos de MAPE, que denotaremos como MAPE. Calculamos también la medida Error de Predicción Cuadrático Medio (MSPE, *Mean Squared Prediction Error*) definida como el promedio de los residuos al cuadrado, es decir, $\text{MSPE} = (1/100)\sum_{i=1}^{100}(y_i - \widehat{y}_i)^2$. Nuevamente, calculamos el promedio sobre los $N_{rep} = 50$ valores de MSPE obteniendo así un único resultado final denotado como MSPE. Además, con el objetivo de medir la selección de variables, en cada iteración calculamos el número de variables seleccionadas como importantes y que deberían efectivamente incluirse en el modelo. Al promediar sobre las $N_{rep} = 50$ replicaciones, obtenemos una medida que llamaremos AV.SIZE, mientras que si calculamos la mediana sobre las $N_{rep} = 50$ replicaciones obtendremos otra medida que denotaremos como MED.SIZE.

Todas estas medidas fueron calculadas sobre la propuesta de estimación y selección de variables desarrollada en esta tesis, que denotaremos como el estimador PENALIZADO, pero así también y a modo de comparación, sobre el estimador que no selecciona variables en ninguna de las dos componentes (que sería equivalente al estimador propuesto en Boente y Martínez, 2023, considerando como $\rho-$función a $\rho(t) = t^2$), que denotaremos como NO PENALIZADO. En este caso, es claro que AV.SIZE y MED.SIZE son iguales a 13. Los resultados de las medidas de calidad de estimación (medidos en términos de MAPE y de MSPE) y de complejidad del modelo (medido en términos de AV.SIZE y MED.SIZE) para el estimador PENALIZADO y el NO PENALIZADO se encuentran en el Cuadro 6.1. Además, se encuentran entre paréntesis los desvíos estándar calculados sobre los 50 valores de MAPE y MSPE.

Estimador	MAPE	MSPE	AV.SIZE	MED.SIZE
NO PENALIZADO	0.4179 (0.0565)	0.9064 (0.0680)	13.00	13.00
PENALIZADO	0.4318 (0.0563)	0.9012 (0.0479)	5.04	5.00

Cuadro 6.1: Valores de MAPE , MSPE, AV.SIZE y MED.SIZE obtenidos tanto para la propuesta de estimación que selecciona variables, como para el estimador NO PENALIZADO. Entre paréntesis se computan los desvíos estándar para la medidas MAPE y MSPE.

El Cuadro 6.1 permite observar que ambos estimadores presentan valores de MAPE y MSPE promedios muy similares, lo que sugiere una capacidad de predicción comparable. Esta similitud se refuerza al considerar los desvíos estándar. Sin embargo, al analizar los valores de AV.SIZE y MED.SIZE obtenidos con el estimador

propuesto en este trabajo de tesis, las diferencias entre ambos procedimientos se vuelven muy notables. Mientras que el estimador NO PENALIZADO considera la totalidad de las 13 variables disponibles, el estimador propuesto reduce sustancialmente la dimensión del modelo, con un valor de AV.SIZE igual a 5.04 y una mediana de 5.00. Esto implica que aproximadamente 5 covariables son identificadas como relevantes y, por consiguiente, aproximadamente ocho de las trece variables disponibles son consideradas como irrelevantes. En conjunto, estos resultados indican que el estimador PENALIZADO logra una capacidad predictiva equiparable a la del estimador que no penaliza, pero con la ventaja adicional de reducir significativamente la cantidad de variables seleccionadas. Esta característica contribuye a la interpretación y parsimonia del modelo resultante, sin sacrificar precisión en la predicción.

Con el objetivo de comparar gráficamente la distribución de los errores de predicción entre ambos estimadores, la Figura 6.1 muestra los boxplots ajustados paralelos de las medida MAPE (izquierda), y de la medida MSPE (derecha) para ambos estimadores. En el caso del MAPE, las distribuciones resultan bastante similares: ambos presentan cajas de tamaño comparable, lo que indica una variabilidad semejante en los valores centrales. Se observan tres valores atípicos en el estimador PENALIZADO y uno en el estimador NO PENALIZADO. La mediana del estimador PENALIZADO es levemente superior, aunque se mantiene dentro del rango intercuartil del estimador NO PENALIZADO, lo que refuerza la idea de una diferencia poco significativa. En cuanto al MSPE, se destaca la menor dispersión observada en el estimador PENALIZADO, evidenciada por un boxplot considerablemente más angosto en comparación con el del estimador NO PENALIZADO.

Es interesante observar cuáles fueron las variables más veces incluidas en el modelo por el estimador PENALIZADO. Para ello, en el Cuadro 6.2 se observa la información de la proporción de veces que se selecciona cada una de las variables en la parte lineal, mientras que en el Cuadro 6.3 se observan las frecuencias con las que son seleccionadas como importantes las covariables incluidas en la componente aditiva del modelo.

SEX	SMOK1	SMOK2	BMI	VIT1	VIT2	CAL	FAT	ALC	BETA
0.26	0.10	**0.84**	**0.98**	0.48	**0.96**	0.10	0.30	0.02	**0.80**

Cuadro 6.2: Frecuencia con la que las variables incluidas en la componente lineal han sido consideradas como significativas para el estimador. Se indica con negrita aquellas que superan el umbral de 0.5.

En el Cuadro 6.2, considerando un umbral de corte 0.5, las variables de la componente lineal identificadas como relevantes en al menos la mitad de las repeticiones son SMOK2, BMI, VIT2 y BETADIET. Por su parte, la variable VIT1

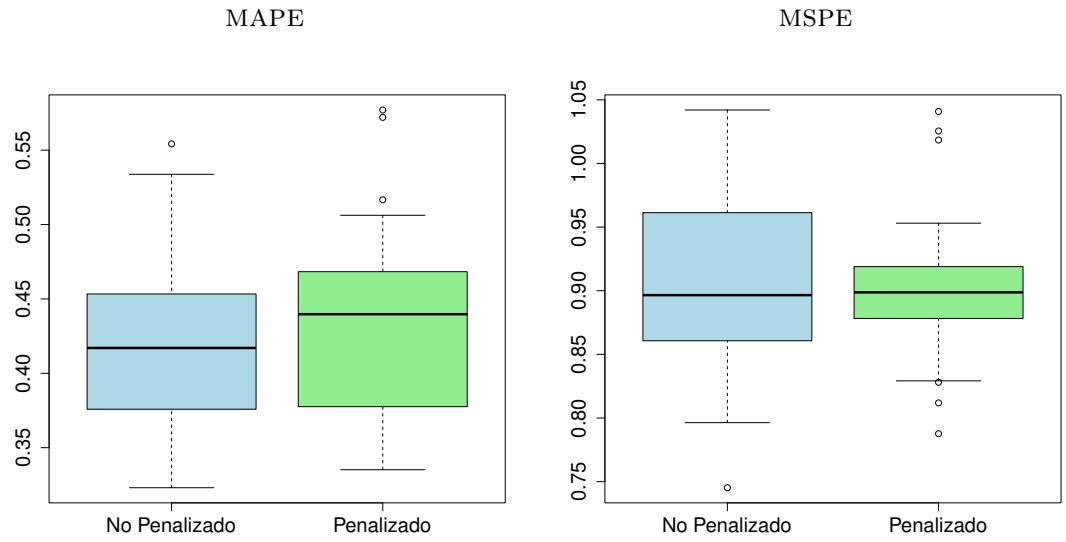

Figura 6.1: Boxplots ajustados de las medidas MAPE a la izquierda y MSPE a la derecha para los estimadores NO PENALIZADO y PENALIZADO.

fue seleccionada en el 48 % de las ocasiones, valor que se encuentra por debajo del umbral considerado.

AGE	CHOL	FIBER
0.00	0.12	0.00

Cuadro 6.3: Frecuencia con la que las variables incluidas en la componente aditiva han sido consideradas como significativas para el estimador.

En relación con el Cuadro 6.3, se observa con claridad que las variables incluidas en el modelo en la parte aditiva son consideradas como irrelevantes para el estimador propuesto. En efecto, tanto AGE como FIBER fueron sistemáticamente descartadas en la totalidad de las 50 replicaciones realizadas, lo cual sugiere de manera consistente que no presentan una contribución significativa al modelo. Por su parte, la variable CHOL fue seleccionada únicamente en 6 ocasiones, lo que representa un 12 % de las repeticiones, muy por debajo del umbral del 50 % utilizado como criterio de inclusión. En conjunto, estos resultados refuerzan la utilidad del procedimiento propuesto para identificar de forma efectiva aquellas variables cuya contribución puede ser despreciada sin pérdida de capacidad predictiva.

En relación con los parámetros de penalización, la Figura 6.2 presenta los boxplots e histogramas de los valores de los parámetros auxiliares λ_1 y λ_2 seleccionados mediante el criterio de BIC definido en la Sección 4.2.

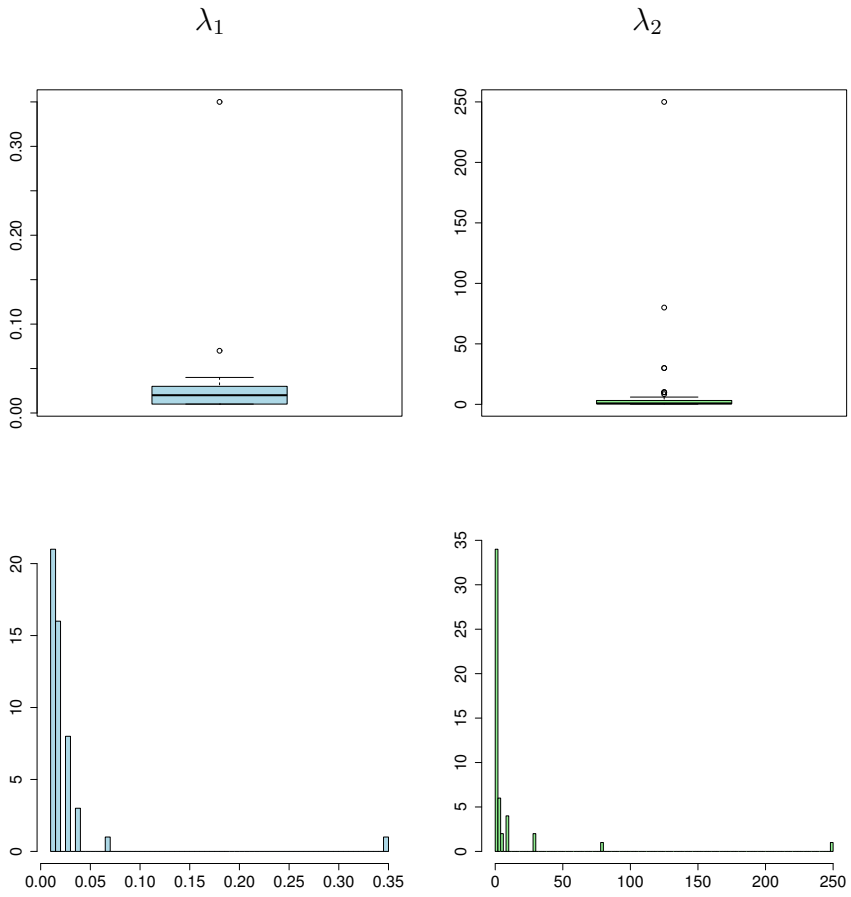

Figura 6.2: Boxplots e histogramas de los valores de λ_1 y λ_2 seleccionados por el criterio de BIC.

Se observa que los valores de λ_1 elegidos suelen ser pequeños, en la mayoría de los casos no superan el valor de 0.05, aunque en un caso atípico alcanza un valor de 0.35. De manera similar, los valores de λ_2 seleccionados también son generalmente bajos, aunque en dos ocasiones superan el valor 50.

Por último, la Figura 6.3 presenta un histograma que resume la distribución de los valores del parámetro k obtenidos a lo largo de las 50 replicaciones. Recordemos que este parámetro controla la cantidad de funciones spline incluidas en las bases utilizadas para representar el efecto de la parte aditiva del modelo. En este gráfico se puede observar que en 47 de las 50 repeticiones se seleccionó un valor de $k = 4$, mientras que en las tres restantes se obtuvo $k = 5$. Esta concentración tan marcada indica la notable estabilidad en la elección de este parámetro de la propuesta de estimación y selección de variables de esta tesis.

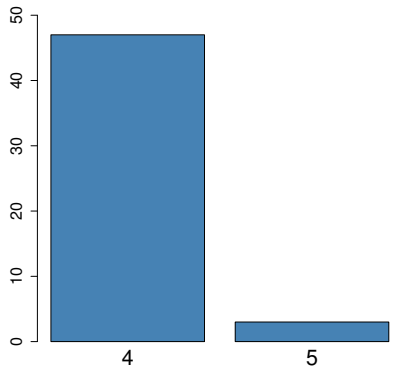

Figura 6.3: Gráfico de barras de los valores de k obtenidos.

6.1. Incorporación de covariables

Para estudiar un poco más en profundidad la capacidad de selección de variables y la calidad de estimación de la propuesta, siguiendo las ideas de Guo et al. (2013) para la componente lineal, en esta sección agregaremos a las covariables existentes, variables no significativas tanto en la componente lineal como en la componente aditiva del modelo. Más específicamente, incluiremos 10 variables uniformes en el intervalo $(0, 1)$ en la componente lineal y 3 en la parte aditiva, lo que representa aumentar al doble la cantidad de covariables que originalmente se consideraron en el modelo. De esta forma, el vector $\mathbf{Z}$ del modelo estará conformado por las 10 variables incluidas originalmente y las 10 variables adicionales, teniendo así $\mathbf{Z} \in \mathbb{R}^{20}$, mientras que el vector $\mathbf{X}$ del modelo estará conformado por las 3 variables incluidas originalmente y las 3 variables adicionales, teniendo así $\mathbf{X} \in \mathbb{R}^{6}$.

En este nuevo escenario, repetimos el análisis anterior junto con todas las medidas mencionadas: MAPE, MSPE para medir la calidad de predicción; y AV.SIZE y MED.SIZE para cuantificar la complejidad del modelo. Además, agregamos la medida FP.PROM que cuenta el promedio, entre las $N_{rep} = 50$ replicaciones, de la cantidad de variables que son incorrectamente seleccionadas como relevantes entre las 13 no significativas que fueron agregadas artificialmente al modelo. Esta información se encuentra disponible en el Cuadro 6.4.

Estimador	mape	mspe	av.size	med.size	fp.prom
no penalizado	0.6654 (0.0546)	1.0472 (0.1167)	25.84	26.00	12.86
penalizado	0.6189 (0.0658)	0.9361 (0.0549)	5.34	5.00	0.62

Cuadro 6.4: Valores de mape , mspe, av.size, med.size y fp.prom obtenidos, tanto para la propuesta de estimación que selecciona variables como para el estimador no penalizado, en el escenario en que se incorporan 13 variables irrelevantes al modelo. Entre paréntesis se computan los desvíos estándar de las medidas mape y mspe.

En el Cuadro 6.4 se observa, nuevamente, que ambos estimadores (penalizado y no penalizado) presentan valores de mape y mspe bastante similares. Sin embargo, en esta ocasión, el estimador penalizado exhibe valores ligeramente inferiores, lo que sugiere una capacidad predictiva comparable, e incluso marginalmente mejor, que la del estimador que no realiza el proceso de penalización. Al incorporar 13 variables que, por construcción sabemos irrelevantes, tanto en la parte lineal como en la aditiva, se evidencia una diferencia clave entre los métodos. El estimador no penalizado incluye, naturalmente, estas variables como relevantes, de hecho, fp.prom = 12.86, mientras que el estimador propuesto selecciona en promedio 0.62 de estas variables irrelevantes, mostrando su gran capacidad de descarte. Por otro lado, el estimador penalizado sigue seleccionando en promedio alrededor de 5 variables como significativas.

Al igual que antes, para observar el comportamiento de las medidas de calidad de la predicción más en detalle, en la Figura 6.4 se presentan los boxplots ajustados para los 50 valores de mape y los 50 valores de mspe obtenidos por el estimador penalizado y el no penalizado.

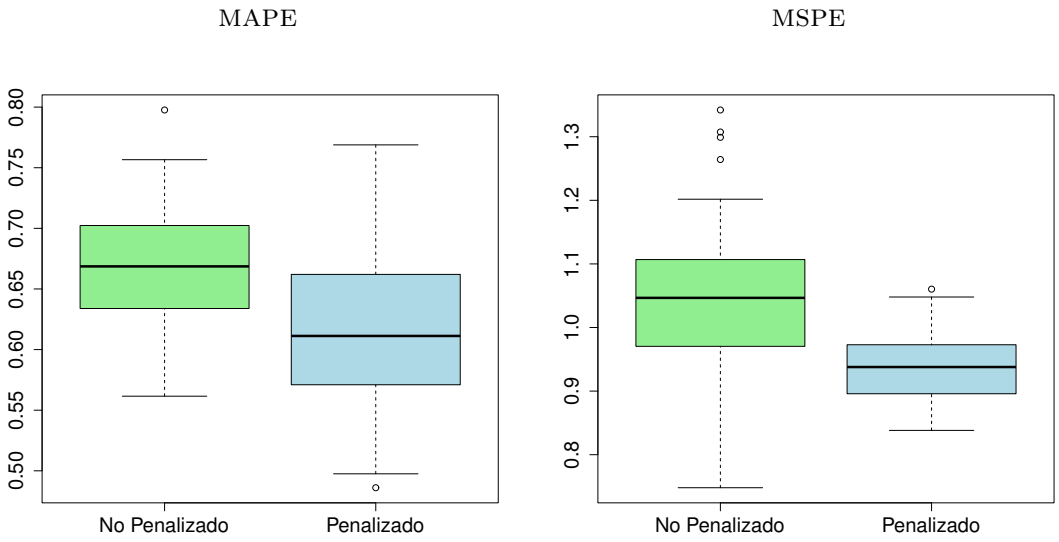

Figura 6.4: Boxplots ajustados de las medidas MAPE a la izquierda y MSPE a la derecha para los estimadores NO PENALIZADO y PENALIZADO cuando se incorporan 13 variables no significativas.

En la Figura 6.4 se observa que los valores de MAPE son considerablemente menores para el estimador PENALIZADO en comparación con el estimador que no selecciona variables. Este comportamiento difiere de lo observado anteriormente, cuando no se incluían las variables irrelevantes. En cuanto a la medida MSPE, se observa que la caja del estimador PENALIZADO se encuentra por debajo de la del estimador NO PENALIZADO y que la caja es más angosta, indicando así un mejor rendimiento predictivo del estimador PENALIZADO. Estos resultados sugieren que, además de una excelente capacidad de selección de variables, el estimador PENALIZADO también es superior en términos predictivos.

Resulta interesante identificar qué variables son ahora consideradas relevantes por el estimador PENALIZADO, habiendo incluido covariables ficticias. En los Cuadros 6.5 y 6.6, vemos que, entre las variables provenientes del conjunto de datos, el estimador sigue considerando las mismas variables que antes como relevantes: SMOK2, BMI, VIT2, y BETADIET. Más aún, se puede observar que los valores de las proporciones variaron muy poco a pesar de analizar un modelo con el doble de covariables de las que tenía originalmente.

SEX	SMOK1	SMOK2	BMI	VIT1	VIT2	CAL	FAT	ALC	BETA
0.30	0.12	**0.68**	**0.88**	0.48	**0.88**	0.12	0.34	0.02	**0.86**

Cuadro 6.5: Frecuencia con la que las variables originales incluidas en la componente lineal han sido consideradas como significativas para el estimador cuando incorporamos al modelo las 13 variables irrelevantes.

AGE	CHOL	FIBER
0.00	0.06	0.00

Cuadro 6.6: Frecuencia con la que las variables originales incluidas en la componente aditiva han sido consideradas como significativas para el estimador cuando incorporamos al modelo las 13 variables irrelevantes.

En el Cuadro 6.7 se presenta la proporción de veces que fueron seleccionadas, a lo largo de las 50 replicaciones, las variables que ingresan al modelo en la parte lineal y que, por construcción sabemos son irrelevantes, mientras que en el Cuadro 6.8 se presenta la proporción de veces que fueron seleccionadas las variables artificiales en la componente aditiva.

V1	V2	V3	V4	V5	V6	V7	V8	V9	V10
0.00	0.04	0.26	0.08	0.02	0.02	0.08	0.08	0.02	0.00

Cuadro 6.7: Proporción de veces que es elegida cada una de las variables ficticias agregadas en la parte lineal cuando incorporamos al modelo las 13 variables irrelevantes.

V11	V12	V13
0.00	0.00	0.00

Cuadro 6.8: Proporción de veces que es elegida cada una de las variables ficticias agregadas en la parte aditiva cuando incorporamos al modelo las 13 variables irrelevantes.

Los resultados muestran un gran desempeño del estimador propuesto. En el Cuadro 6.7 se puede ver que de las 10 variables irrelevantes consideradas en la parte lineal, dos de ellas nunca fueron seleccionadas en ninguna de las replicaciones. Otras tres variables fueron elegidas únicamente una vez, una variable fue seleccionada en dos ocasiones, y tres variables fueron consideradas relevantes en cuatro de las 50 replicaciones. La única excepción destacable es una variable que fue seleccionada en el 26 % de los casos. Sin embargo, este valor continúa por debajo del umbral del 50 % utilizado. Por su parte, en el Cuadro 6.8 se observa un comportamiento aún más contundente: ninguna de las tres variables artificiales fueron seleccionadas como relevantes en ninguna de las 50 replicaciones realizadas.

Este resultado sugiere que el procedimiento de selección aplicado en la componente aditiva posee una mayor capacidad para excluir variables espurias, al menos en este conjunto de datos.

Con respecto a la selección de los parámetros auxiliares de penalización, en la Figura 6.5 se pueden observar los histogramas y boxplots de los λ_1 y λ_2 seleccionados por el criterio de BIC propuesto en la Sección 4.2 para el modelo que incluye 13 covariables artificiales. Podemos observar que, en lo que respecta a λ_1 se siguen obteniendo valores pequeños, aunque en este caso fueron seleccionados algunos valores aislados más grandes. En cuanto a λ_2, los valores seleccionados por el criterio fueron, en varias ocasiones, más grandes que en el modelo que sólo tenía las variables originales. De hecho, en siete ocasiones, fueron seleccionados valores mayores a 50, llegando incluso en tres de ellas a 250.

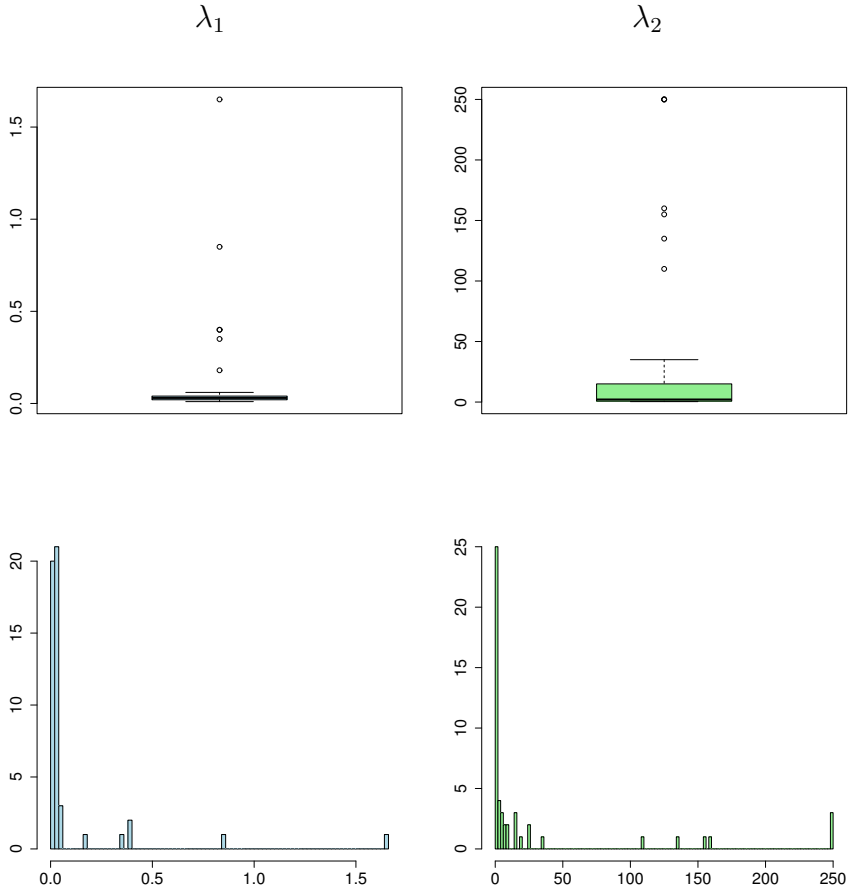

Figura 6.5: Boxplots e histogramas de los parámetros de penalización λ_1 y λ_2 seleccionados por el criterio de BIC cuando al modelo se le incorporan 13 variables irrelevantes.

Por último, en cuanto a la selección del parámetro de suavizado k, en la Figura 6.6 puede verse que, según el criterio de BIC utilizado para este parámetro y definido en la Sección 4.3, el valor $k = 4$ fue seleccionado en 44 oportunidades, mientras que en 5 oportunidades fue seleccionado el valor $k = 5$ y, una única vez, el valor $k = 6$. Esto muestra, nuevamente, la consistencia del estimador en seleccionar el valor $k = 4$ en las sucesivas iteraciones.

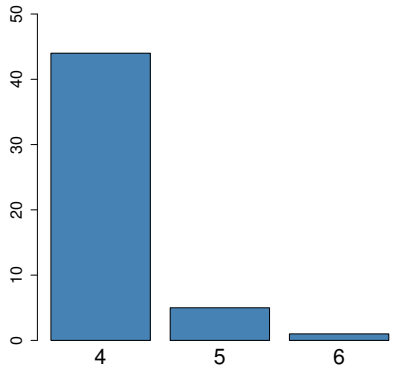

Figura 6.6: Gráfico de barras de los valores de k obtenidos.

En conclusión, el estudio realizado sobre el conjunto de datos reales permitió evaluar el desempeño del procedimiento propuesto en un escenario realista, donde se supuso un modelo parcialmente lineal aditivo y se evaluó la presencia de posibles variables irrelevantes. Se analizaron tanto las capacidades predictivas del modelo como su comportamiento en términos de selección de variables. Los resultados obtenidos mostraron que el estimador propuesto es capaz de descartar de manera efectiva las variables irrelevantes, tanto de la parte lineal como de la parte aditiva del modelo. En particular, la incorporación de variables artificiales irrelevantes permitió verificar la capacidad del estimador de descartar variables, al exhibir una baja tasa de inclusión de dichas covariables. Asimismo, la comparación entre la versión penalizada del modelo y su contraparte sin penalización mostró ventajas claras en términos de parsimonia, manteniendo un buen desempeño predictivo.

Capítulo 7

Conclusiones

En este trabajo se propuso un procedimiento para la estimación y selección simultánea de variables en modelos de regresión parcialmente lineales aditivos (PLAM). Para ello, se utilizó una aproximación mediante $B-$splines para modelar la componente no paramétrica y capturar estructuras no lineales. Asimismo, se consideró un procedimiento de regularización que penaliza ambas componentes del modelo. Con el fin de reducir la búsqueda de los parámetros de penalización a un espacio de dimensión menor, se incorporó un enfoque adaptativo. Finalmente, se introdujeron criterios de tipo BIC para la selección automática tanto de los parámetros de penalización como de suavidad. La propuesta metodlógica incluye, además, el algoritmo necesario para su implementación computacional.

En cuanto al análisis del desempeño del método, en una primera instancia se realizó un estudio de simulación. Se consideraron doce escenarios distintos, variando la cantidad de covariables incluidas en las componentes lineal y aditiva del modelo, así como el número de covariables relevantes, y tres tamaños de muestra diferentes. El comportamiento de la propuesta se evaluó mediante medidas de calidad de la selección de variables y de la estimación, tanto en la parte lineal como en la aditiva. Los resultados indicaron que el procedimiento logra una estimación precisa de las componentes del modelo y, al mismo tiempo, identifica de forma eficaz las covariables relevantes, evitando la inclusión de variables irrelevantes.

El análisis de un conjunto de datos reales permitió ilustrar la aplicabilidad y ventajas del método propuesto en un contexto práctico. En particular, se compararon los resultados obtenidos con los del estimador que no selecciona variables, evidenciándose que la utilización de una propuesta que además de estimar selecciona variables mejora la interpretabilidad y parsimonia del modelo sin comprometer su capacidad predictiva. Además, se incorporaron covariables ficticias tanto en la componente lineal como en la componente aditiva del modelo, lo cual permitió

evidenciar el buen desempeño del método incluso en escenarios con un número mayor de covariables.

En conclusión, el procedimiento propuesto en esta tesis demostró ser eficaz para la estimación y selección de variables en modelos PLAM. Constituye una herramienta flexible y eficiente, adecuada para el análisis de datos en contextos con un elevado número de variables, donde es necesario modelar tanto efectos lineales como relaciones más generales desconocidas de forma simultánea.

Bibliografía

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. *Second International Symposium on Information Theory*, 267–281.

Banerjee, S., y Ghosal, S. (2014). Bayesian variable selection in generalized additive partial linear models. *Stat*, *3*(1), 363–378.

Boente, G., y Martínez, A. (2023). A robust spline approach in partially linear additive models. *Computational Statistics and Data Analysis*, *178*, 107611. doi: 10.1016/j.csda.2023.107611

Boente, G., y Martínez, A. (2024). Robust variable selection for partially linear additive models. *Statistics and Computing*, *34*(201), 1–18. doi: 10.1007/s11222-024-10520-7

Breiman, L. (1996). Bagging predictors. *Machine Learning*, *24*(2), 123–140. doi: 10.1007/BF00058655

Buja, A., Hastie, T., y Tibshirani, R. (1989). Linear smoothers and additive models. *The Annals of Statistics*, 453–510.

Bunea, F. (2004). Consistent covariate selection and post model selection inference in semiparametric regression. *The Annals of Statistics*, *32*(2), 898–927.

Bunea, F., y Wegkamp, M. H. (2004). Two-stage model selection procedures in partially linear regression. *Canadian Journal of Statistics*, *32*(2), 105–118.

Chebi, G. (2019). *M-estimadores penalizados para regresión logística* (Tesis Doctoral, Universidad de Buenos Aires, Facultad de Ciencias Exactas y Naturales). Descargado de `https://hdl.handle.net/20.500.12110/tesis_n6891_Chebi`

Draper, N. R. (1998). *Applied regression analysis* (3rd ed.). New York: Wiley.

Du, P., Cheng, G., y Liang, H. (2012). Semiparametric regression models with additive nonparametric components and high dimensional parametric components. *Computational Statistics & Data Analysis*, *56*(6), 2006–2017.

Efron, B., Hastie, T., Johnstone, R., y Tibshirani, R. (2004). Least angle regression. *The Annals of statistics*, *32*(2), 407–499.

Fan, J., y Li, R. (2001). Variable selection via nonconcave penalized likelihood estimation. *Journal of the American Statistical Association*, *96*(456), 1348–1360. doi: 10.1198/016214501753382273

Fan, J., y Li, R. (2002). Variable selection for cox's proportional hazards model and frailty model. *The Annals of Statistics*, *30*(1), 74–99.

Fan, J., y Li, R. (2004). New estimation and model selection procedures for semiparametric modeling in longitudinal data analysis. *Journal of the American Statistical Association*, *99*(467), 710–723.

Fan, J., y Lv, J. (2011). Nonconcave penalized likelihood with np-dimensionality. *IEEE Transactions on Information Theory*, *57*(8), 5467–5484.

Feng, M., Chen, Z., y Cheng, X. (2018). Two-step variable selection in partially linear additive models with time series data. *Communications in Statistics - Simulation and Computation*, *47*, 661–671.

Foster, D. P., y George, E. I. (1994). The risk inflation criterion for multiple regression. *The Annals of Statistics*, *22*(4), 1947–1975.

Frank, I., y Friedman, J. (1993). A statistical view of some chemometrics regression tools. *Technometrics*, *35*(2), 109–148. doi: 10.1080/00401706.1993.10485254

Gujarati, D. N. (2009). *Basic econometrics* (5th ed.). New York: McGraw-Hill.

Guo, J., Tang, M., Tian, M., y Zhu, K. (2013). Variable selection in high-dimensional partially linear additive models for composite quantile regression. *Computational Statistics & Data Analysis*, *65*, 56–67.

Hastie, T., y Tibshirani, R. (1986). Generalized additive models. *Statistical Science*, *1*(3), 297–310. doi: 10.1214/ss/1177013604

Hastie, T., Tibshirani, R., y Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). New York: Springer.

Hastie, T., Tibshirani, R., y Wainwright, M. (2015). *Statistical learning with sparsity: The lasso and generalizations*. CRC Press.

Hastie, T. J. (2017). *Computer age statistical inference: Algorithms, evidence, and data science*. Cambridge University Press.

Hastie, T. J., y Tibshirani, R. J. (1990). Generalized additive models (monographs on statistics and applied probability 43). *London: Chapman&Hall/CRC*.

He, X., Zhu, Z. Y., y Fung, W. K. (2002). Estimation in a semiparametric model for longitudinal data with unspecified dependence structure. *Biometrika*, *89*, 579–590. doi: 10.1093/biomet/89.3.579

Hocking, R. R. (1976). The analysis and selection of variables in linear regression. *Biometrics*, *32*(1), 1–49. doi: 10.2307/2529336

Hoerl, A. E., y Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, *12*(1), 55–67.

Hu, Y., Zhao, K., y Lian, H. (2015). Bayesian quantile regression for partially linear additive models. *Statistics and Computing*, *25*, 651–668.

Hubert, M., y Vandervieren, E. (2008). An adjusted boxplot for skewed distributions. *Computational Statistics and Data Analysis*, *52*(12), 5186-5201. doi: https://doi.org/10.1016/j.csda.2007.11.008

James, G., Witten, D., Hastie, T., y Tibshirani, R. (2013). *An introduction to statistical learning: With applications in r*. New York: Springer. doi: 10.1007/978-1-4614-7138-7

Jeong, S., Park, T., y Van Dyk, D. A. (2022). Bayesian model selection in additive partial linear models via locally adaptive splines. *Journal of Computational and Graphical Statistics*, *31*(2), 324–336.

Kazemi, M., Shahsavani, D., y Arashi, M. (2018). Variable selection and structure identification for ultrahigh-dimensional partially linear additive models with application to cardiomyopathy microarray data. *Statistics, Optimization & Information Computing*, *6*(3), 373–382.

Kutner, M. H., Nachtsheim, C. J., Neter, J., y Li, W. (2004). *Applied linear statistical models* (5th ed.). Boston: McGraw-Hill/Irwin.

Li, R., y Liang, H. (2008). Variable selection in semiparametric regression modeling. *Annals of statistics*, *36*(1), 261.

Liang, H., y Li, R. (2009). Variable selection for partially linear models with measurement errors. *Journal of the American Statistical Association*, *104*(485), 234–248.

Linton, O., y Nielsen, H. (1995). Nonparametric estimation of additive models. *Journal of Multivariate Analysis*, *55*(1), 75–97. doi: 10.1006/jmva.1995.1024

Liu, J., Lou, L., y Li, R. (2018). Variable selection for partially linear models via partial correlation. *Journal of multivariate analysis*, *167*, 418–434.

Liu, X., Wang, L., y Liang, H. (2011). Estimation and variable selection for semiparametric additive partial linear models (ss-09-140). *Statistica Sinica*, *21*(3), 1225.

Lv, J., Yang, H., y Guo, C. (2017). Variable selection in partially linear additive models for modal regression. *Communications in Statistics-Simulation and Computation*, *46*(7), 5646–5665.

Meinshausen, N., y Bühlmann, P. (2010). Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *72*(4), 417–473.

Ni, X., Zhang, H. H., y Zhang, D. (2009). Automatic model selection for partially linear models. *Journal of multivariate Analysis*, *100*(9), 2100–2111.

Nierenberg, D. W., Stukel, T. A., Baron, J. A., Dain, B. J., Greenberg, E. R., y Group, S. C. P. S. (1989). Determinants of plasma levels of beta-carotene and retinol. *American Journal of Epidemiology*, *130*(3), 511–521.

Opsomer, J. D., y Ruppert, D. (1997). Fitting a bivariate additive model by local polynomial regression. *The Annals of Statistics*, *25*(1), 186–211.

Opsomer, J. D., y Ruppert, D. (1999). A root-n consistent backfitting estimator for semiparametric additive modeling. *Journal of Computational and Graphical Statistics*, *8*(4), 715–732.

Schoenberg, I. J. (1946). Contributions to the problem of approximation of equidistant data by analytic functions. *Quarterly of Applied Mathematics*, *4*, 45–99 and 112–141.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*(2), 461–464. doi: 10.1214/aos/1176344136

Stone, C. J. (1985). Additive regression and other nonparametric models. *The annals of Statistics*, 689–705.

Sun, Y., y Genton, M. G. (2011). Functional boxplots. *Journal of Computational and Graphical Statistics*, *20*, 316-334.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *58*(1), 267–288. doi: 10.1111/j.2517-6161.1996.tb02080.x

Tjøstheim, T., y Auestad, B. (1994). Nonparametric regression and additive models. *Statistical Science*, *9*(4), 421–448. doi: 10.1214/ss/1177011074

Wang, H., y Leng, C. (2008). A note on adaptive group lasso. *Computational Statistics and Data Analysis*, *52*(12), 5277–5286.

Wei, F. (2012). Group selection in high-dimensional partially linear additive models. *Brazilian Journal of Probability and Statistics*, *26*, 219–243.

Wooldridge, J. M. (2010). *Econometric analysis of cross section and panel data* (2nd ed.). Cambridge: MIT Press.

Xie, H., y Huang, J. (2009). Scad-penalized regression in high-dimensional partially linear models. *Ann Statist*, *37*, 673–696.

Yuan, M., y Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *68*(1), 49–67.

Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, *38*(2), 894–936. doi: 10.1214/09-AOS731

Zhao, P., y Xue, L. (2009). Variable selection for semiparametric varying coefficient partially linear models. *Statistics & Probability Letters*, *79*(20), 2148–2157.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, *101*(476), 1418–1429. doi: 10.1198/016214506000000735

Zou, H., y Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *67*(2), 301–320. doi: 10.1111/j.1467-9868.2005.00503.x