Facultad de Informática

Selección de variables: Una revisión de métodos existentes

Máster en Técnicas Estadísticas

Autora:

Ana González Vidal

Director:

Germán Aneiros Pérez

Trabajo Fin de Máster

Curso: 2014-2015

El presente documento recoge bajo el título *Selección de variables: Una revisión de métodos existentes*, el trabajo realizado por Dña. Ana González Vidal como Trabajo Fin de Máster del máster interuniversitario en Técnicas Estadísticas bajo la dirección de D. Germán Aneiros Pérez, profesor titular del departamento de Matemáticas de la Universidad de A Coruña.

Fdo: Ana González Vidal

Fdo: Germán Aneiros Pérez

Agradecimientos

A mi tutor, Germán Aneiros Pérez, así como a todos los profesores y compañeros del máster.

A mi familia y amigos. A todos ellos, muchas gracias.

Resumen

El tópico a tratar en este TFM es el de la selección de variables explicativas a ser incorporadas en un modelo de regresión lineal. Se considerará tanto el caso en que el total de covariables, p, es inferior al tamaño muestral, n, como el caso correspondiente a p mucho mayor que n. En primer lugar, se presentarán diferentes métodos de selección de variables así como las propiedades que cumplen cada uno de ellos. Posteriormente se comparará, a través de un estudio de simulación, el comportamiento de dichos métodos de selección de variables. El TFM finalizará aplicando dichos métodos a algún conjunto de datos reales. En la implementación (simulación y aplicación a datos reales) se utilizarán rutinas ya existentes en distintos paquetes de R.

Índice general

Capítulo 1

Introducción

Este primer capítulo estará dedicado a estudiar de forma breve en que consiste un modelo de regresión. Se dividirá en diferentes secciones de la siguiente manera; en la Sección 1.1 comenzaremos introduciendo de forma breve el modelo de regresión lineal, tanto el modelo de regresión lineal simple como múltiple. A continuación en la Sección 1.2 abordaremos el problema de la estimación de los parámetros del modelo mediante el método de mínimos cuadrados. Seguidamente en la Sección 1.3 motivaremos la necesidad de la selección de variables en un modelo de regresión y por último en la Sección 1.4 introduciremos notación general que usaremos a lo largo del trabajo.

1.1. Modelo de regresión lineal

Para representar la dependencia de una variable Y (variable dependiente, variable respuesta) con respecto a otra variable X (variable independiente, variable explicativa), se utilizan los modelos de regresión. Dichos modelos se diseñan con dos objetivos:

- Describir la forma de dependencia, es decir, conocer de que modo la variable Y depende de la variable X.
- Por otra parte, una vez que tenemos construido el modelo, también nos sirve para realizar predicciones de un valor de la variable Y cuando se conoce el valor de la

variable X.

Supongamos que tenemos una variable respuesta y una variable explicativa, entonces el modelo de regresión lineal se define de la siguiente manera:

$$Y = \beta_1 X + \varepsilon$$

siendo Y la variable respuesta y X la variable explicativa. El coeficiente β_1 es el parámetro que acompaña a la variable explicativa y que queremos estimar y ε el error aleatorio.

Por comodidad, hemos asumido que las variables de todos los modelos incluidos en este trabajo tienen media cero; de este modo, el término independiente de tales modelos es nulo y no necesitamos ponerlo.

Este modelo de regresión lineal simple, se puede extender a situaciones más complejas en las cuales hay más de una variable explicativa. Supongamos que tenemos una variable respuesta Y y una colección de variables explicativas $X_1, X_2, \ldots, X_p$. Para obtener un modelo de regresión múltiple, basta considerar una combinación lineal de las variables explicativas de la siguiente manera:

$$Y = \beta_1 X_1 + \ldots + \beta_p X_p + \varepsilon.$$

Los coeficientes $\beta_1, \ldots, \beta_p$ son los parámetros desconocidos que acompañan a las variables explicativas y que van a ser estimados y ε el error aleatorio, que suponemos que cumple las hipótesis de homocedasticidad y media nula (clásicamente, también se asume normalidad e independencia).

1.2. Estimación de los parámetros mediante el método de mínimos cuadrados

En esta sección abordaremos el problema de la estimación del vector de parámetros desconocido del modelo: $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$. La estimación se realizará mediante el método

de mínimos cuadrados. Partimos del modelo

$$y_i = x_{i1}\beta_1 + \ldots + x_{ip}\beta_p + \varepsilon_i$$

para $i = 1, \ldots, n$, que escribiremos también como $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, donde

$$\boldsymbol{Y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \quad \boldsymbol{X} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{i1} & x_{i2} & \cdots & x_{ip} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix} \text{ y } \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

siendo $\boldsymbol{X}$ la matriz de diseño de dimensión $n \times p$ con n el tamaño de muestra y p el número de variables explicativas, es decir, una matriz donde la fila i-ésima representa las observaciones del i-ésimo individuo en las p variables explicativas y la columna j-ésima representa las observaciones de la j-ésima variable en los n individuos. Nuestro objetivo es estimar el vector de parámetros $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^{'}$, para lo cual utilizaremos el método de mínimos cuadrados. Así, escogeremos como estimador de $\boldsymbol{\beta}$ aquel $\widehat{\boldsymbol{\beta}}_{\boldsymbol{n}}$ donde se alcance

$$\min_{\beta} \sum_{i=1}^{n} (y_i - \boldsymbol{x_i^{'}}\boldsymbol{\beta})^2$$

siendo y_i la observación de la variable respuesta correspondiente al individuo i-ésimo, es decir el elemento i-ésimo de la variable $\boldsymbol{Y}$ y $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})^{'}$ el vector correspondiente a la fila i-ésima de la matriz de diseño $\boldsymbol{X}$, es decir, las observaciones del i-ésimo individuo en las p variables explicativas.

En notación matricial, el problema de minimización se puede expresar de la siguiente manera:

$$\min_{\beta} (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^{'}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) = \min_{\beta} \phi(\boldsymbol{\beta})$$

1.3. Selección de variables

En esta sección motivaremos la necesidad de seleccionar variables en un modelo de regresión.

En muchas situaciones se dispone de un conjunto grande de posibles variables explicativas, una posible pregunta sería saber si todas las variables deben entrar en el modelo de regresión, y en caso negativo, saber que variables deben entrar y cuales no.

Los métodos de selección de variables se encargan de abordar el problema de construcción o selección del modelo. En general, si se incluyen cada vez más variables en un modelo de regresión, el ajuste a los datos mejora, aumenta la cantidad de parámetros a estimar pero disminuye su precisión individual (mayor varianza) y por tanto la de la función de regresión estimada, se produce un sobreajuste. Por el contrario, si se incluyen menos variables de las necesarias en el modelo, las varianzas se reducen pero los sesgos aumentarán obteniéndose una mala descripción de los datos. Por otra parte, algunas variables predictoras pueden perjudicar la confiabilidad del modelo, especialmente si están correlacionadas con otras. De esta manera, el objetivo de los métodos de selección de variables es buscar un modelo que se ajuste bien a los datos y que a la vez sea posible buscar un equilibrio entre bondad de ajuste y sencillez.

El **método de mínimos cuadrados ordinarios** es una forma común de estimar el vector de parámetros desconocido $\boldsymbol{\beta}$ del modelo de regresión. Presenta un buen comportamiento en el caso de que todos los parámetros del modelo sean relevantes. Hay dos razones por las que el método de mínimos cuadrados podría no ser adecuado para estimar modelos con variables no relevantes.

- **Precisión de la predicción**: Las estimaciones de los parámetros por mínimos cuadrados tienen bajo sesgo pero gran varianza. La precisión de la predicción a veces se puede mejorar mediante la reducción o ajuste a cero de algunos coeficientes.
- **Interpretación**: Con un gran número de predictores a menudo nos gustaría determinar un subconjunto más pequeño que exiba los efectos más fuertes.

Al estimar los parámetros de regresión por el método de mínimos cuadrados ordinarios, puede que alguna de estas estimaciones sea casi cero y por tanto la variable correspondiente a dicho coeficiente tendría muy poca influencia en el modelo, sin embargo, es poco común que estas estimaciones lleguen a tomar exactamente el valor cero, por tanto, no nos sirve como método de selección de variables. De este modo, necesitaremos de otros métodos para seleccionar variables, como pueden ser los siguientes:

- **Algoritmos**: Entre los algoritmos para seleccionar variables podemos destacar los siguientes, que explicaremos detalladamente en el siguiente capítulo.
  - **Métodos Forward**: Consiste en la selección de variables hacia adelante.
  - **Métodos Backward**: Se basa en la eliminación de variables hacia atrás.
  - **Métodos Stepwise**: Engloban una serie de procedimientos de selección automática de variables significativas, basados en la inclusión o exclusión de las mismas en el modelo de una manera secuencial.
  - **Best subset**: Se basa en encontrar un subconjunto de variables que proporcione el mejor modelo.
- **Métodos de mínimos cuadrados penalizados**: Se basan en los mínimos cuadrados ordinarios pero añadiendo una penalización en la función objetivo, para forzar que alguna componente del vector de parámetros $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$ sea cero y de esta manera conseguir estimación de los parámetros y selección de variables conjuntamente.

1.4. Notación General

Esta sección estará dedicada a establecer la notación general que se utilizará a lo largo de este trabajo.

Como habíamos visto en la sección anterior a lo largo de este trabajo se partirá del

siguiente modelo de regresión

$$y_i = x_{i1}\beta_1 + \ldots + x_{ip}\beta_p + \varepsilon_i \quad \text{para } i = 1, \ldots, n$$

que se podía escribir también como $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$.

Denotaremos por $\widehat{\boldsymbol{\beta}}_{\boldsymbol{n}} = (\widehat{\beta}_1, \ldots, \widehat{\beta}_p)'$ al estimador del vector de parámetros $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$.

A continuación vamos a denotar por $\boldsymbol{S} = \{j/\beta_j \neq 0\}$ al conjunto de índices tal que el parámetro correspondiente a ese índice es distinto de cero y por $\widehat{\boldsymbol{S}} = \{j/\widehat{\beta}_j \neq 0\}$ al conjunto de índices tal que el parámetro estimado correspondiente a ese índice es distinto de cero.

Veremos como a lo largo del trabajo se mostrarán determinadas propiedades de los diferentes métodos de selección de variables para las cuales nos será útil dividir el vector de parámetros en dos subvectores de la siguiente manera:

Sea $\boldsymbol{\beta} = (\boldsymbol{\beta_1'}, \boldsymbol{\beta_2'})'$ el vector de parámetros desconocido que queremos estimar. Sin pérdida de generalidad asumiremos que $\boldsymbol{\beta_1} = (\beta_1, \ldots, \beta_{kn})'$ es el vector de parámetros no nulos de dimensión $k_n \times 1$ y $\boldsymbol{\beta_2} = (0, \ldots, 0)'$ el vector de parámetros nulos de dimensión $m_n \times 1$. Aquí $k_n (\leq p)$ es el número de parámetros no nulos y $m_n = p - k_n$ el número de parámetros nulos. Expresaremos por $\widehat{\boldsymbol{\beta}}_{\boldsymbol{n}} = (\widehat{\boldsymbol{\beta}}'_{\boldsymbol{1n}}, \widehat{\boldsymbol{\beta}}'_{\boldsymbol{2n}})'$ los estimadores de los parámetros $\boldsymbol{\beta_1}$ y $\boldsymbol{\beta_2}$ respectivamente.

Capítulo 2

Métodos de selección de variables

En el capítulo anterior hemos motivado la necesidad y la importancia de la selección de variables en modelos de regresión e introducido de forma breve los algoritmos y métodos de selección de variables más usados. Este nuevo capítulo estará dedicado a su descripción. En la Sección 2.1 explicaremos diferentes algoritmos de selección de variables, como los algoritmos stepwise, el algoritmo para la selección del mejor subconjunto, etc. Por último, en la Sección 2.2 indicaremos en que consiste el método de mínimos cuadrados penalizados e introduciremos brevemente alguno de ellos, como por ejemplo el método Lasso, los métodos Bridge y el método SCAD.

2.1. Algoritmos para la selección de variables

La idea de los algoritmos de selección de variables es elegir el mejor modelo en forma secuencial pero incluyendo o excluyendo una sola variable predictora en cada paso de acuerdo a ciertos criterios.

El proceso secuencial termina cuando se satisface una regla de parada establecida.

A continuación se describen tres de los algoritmos más usados.

- **Métodos Forward**: (Selección hacia adelante). Se parte de un modelo muy sencillo y se van agregando términos con algún criterio, hasta que no procede añadir ningún

término más, es decir, en cada etapa se introduce la variable más significativa hasta una cierta regla de parada.

- **Métodos Backward**: (Eliminación hacia atrás). Se parte de un modelo muy complejo, que incorpora todos los efectos que pueden influir en la respuesta, y en cada etapa se elimina la variable menos influyente, hasta que no procede suprimir ningún término más.

- **Métodos Stepwise**: Este procedimiento es una combinación de los dos anteriores. Comienza como el de introducción progresiva, pero en cada etapa se plantea si todas las variables introducidas deben de permanecer en el modelo.

Cuando se aplica este tipo de procedimientos tenemos que tener en cuenta cual será la condición para suprimir o incluir un término. Para ello podemos considerar dos criterios: criterios de significación del término y criterios de ajuste global.

- **Criterios de significación**: En un método **backward** se suprimirá el término que resulte menos significativo, y en un método **forward** se añadirá el término que al añadirlo al modelo resulte más significativo. Un criterio de significación puede ser la significación de cada coeficiente.

- **Criterios globales**: En vez de usar la significación de cada coeficiente, podemos basarnos en un criterio global, una medida global de cada modelo, de modo que tenga en cuenta el ajuste y el exceso de parámetros. Escogeremos el modelo cuya medida global sea mejor. Como criterios destacamos el Criterio de Información de Akaike (AIC) y el Criterio de Información de Bayes (BIC). Se trata de buscar un modelo cuyo AIC o BIC sea pequeño, ya que en ese caso habría una verosimilitud muy grande y pocos parámetros.

Otro de los métodos de selección de variables es el siguiente:

Best subset: Dadas p variables explicativas, este método consiste en formar todos los posibles subconjuntos de variables explicativas y efectuar todas las posibles regresiones,

reteniendo aquella que, de acuerdo con el criterio de bondad de ajuste que hayamos elegido, parezca mejor (coeficiente de determinación, coeficiente de determinación ajustado y coeficiente C_p de Mallows). El inconveniente de este método es el gran volumen de cálculo que es preciso realizar.

Además de estos algoritmos para la selección de variables habíamos propuesto también el método de mínimos cuadrados penalizados como método de selección. Por tanto, en la siguiente sección se explicará en que consiste dicho método y se introducirá de forma breve los métodos de mínimos cuadrados penalizados más destacados, como son el método Lasso, los métodos Bridge y el método SCAD.

2.2. Métodos de mínimos cuadrados penalizados

Los algoritmos estudiados anteriormente, pueden resultar fuertemente inestables o directamente inaplicables cuando el número de variables p es similar o incluso superior al número de observaciones n. Una alternativa son los métodos de regresión penalizada. La idea clave es la penalización, se evita el sobreajuste debido al gran número de variables predictoras imponiendo una penalización o término de penalización, que obligará a que alguna componente del vector de parámetros $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$ sea cero. La elección del parámetro de penalización es fundamental: es necesario un procedimiento que estime el valor de dicho parámetro a partir de los datos. Por tanto, en un intento de seleccionar las variables y de estimar los parámetros de forma automática y simultánea, se propone un enfoque unificado a través de mínimos cuadrados penalizados, que consiste en estimar el vector de parámetros $\boldsymbol{\beta}$, minimizando la siguiente expresión

$$\sum_{i=1}^{n}(y_i - \boldsymbol{x_i'}\boldsymbol{\beta})^2 + \lambda \sum_{j=1}^{p} p_\lambda(|\beta_j|), \tag{2.1}$$

donde p_λ es la función de penalización que será diferente para cada método y λ es el parámetro de penalización. En cuanto al parámetro λ debe ser elegido a través de algún procedimiento basado en los datos muestrales.

De este modo, estimaremos el vector de parámetros $\boldsymbol{\beta}$ como el que minimiza la expresión (2.1) y se denotará como $\widehat{\boldsymbol{\beta}}_{\boldsymbol{n}}$.

Naturalmente si $\lambda = 0$ este estimador se corresponde con el estimador de mínimos cuadrados ordinarios, que es denotado por $\widehat{\boldsymbol{\beta}}_{\boldsymbol{n}}^{(\mathbf{0})}$.

A continuación se proponen tres condiciones deseables que un método de penalización debería cumplir:

- Esparsidad: efectuar selección de variables automáticamente, esto es, tener la potencialidad de fijar coeficientes a cero.
- Continuidad: Ser continuo en los datos para evitar inestabilidad en la predicción.
- Insesgadez: tener bajo sesgo, especialmente para valores grandes de los coeficientes β_j.

A continuación se describen tres métodos de regresión penalizada. El método Lasso, los métodos Bridge y el método SCAD, los cuales difieren en el tipo de penalización (p_λ) utilizada.

2.2.1. Método Lasso

El Método Lasso (*least absolute shrinkage and selection operator*) introducido por Tibshirani (1996) es un método que combina contracción de algunos parámetros hacia cero y selección de variables, imponiendo una restricción o una penalización sobre los coeficientes de regresión. Así siendo $\boldsymbol{\beta}$ el vector de coeficientes, el método Lasso lo estima minimizando el siguiente problema de mínimos cuadrados penalizados

$$\min_{\beta} \sum_{i=1}^{n} (y_i - \boldsymbol{x_i'}\boldsymbol{\beta})^2 \tag{2.2}$$

sujeto a

$$\sum_{j=1}^{p} |\beta_j| \leq t$$

o de forma equivalente minimizando la siguiente expresión

$$\sum_{i=1}^{n}(y_i - \boldsymbol{x_i'}\boldsymbol{\beta})^2 + \lambda \sum_{j=1}^{p} |\beta_j| \tag{2.3}$$

siendo t y λ los parámetros de regularización o de penalización.

Para valores grandes de λ o valores pequeños de t, los coeficientes β_j se contraen hacia cero y alguno de ellos se anula, por eso se dice que Lasso produce estimación de parámetros y selección de variables simultánea.

2.2.2. Método Bridge

Una familia de funciones de penalización muy utilizada es la correspondiente a la norma L_q dada por $\sum_{j=1}^{p} p_\lambda(|\beta_j|) = \sum_{j=1}^{p} |\beta_j|^q$. Los estimadores resultantes en este caso son conocidos como estimadores Bridge. Estos métodos propuestos por Fu (1998) y Frank y Friedman (1993) resuelven el siguiente problema de minimización

$$\min_{\beta} \sum_{i=1}^{n}(y_i - \boldsymbol{x_i'}\boldsymbol{\beta})^2 \tag{2.4}$$

sujeto a

$$\sum_{j=1}^{p} |\beta_j|^\gamma \leq t$$

o de forma equivalente minimizando

$$\sum_{i=1}^{n}(y_i - \boldsymbol{x_i'}\boldsymbol{\beta})^2 + \lambda \sum_{j=1}^{p} |\beta_j|^\gamma \tag{2.5}$$

siendo t y λ los parámetros de regularización o de penalización. Cuando $\gamma = 1$ se trata del método Lasso explicado anteriormente y cuando $\gamma = 2$ este método se conoce como método Ridge.

2.2.3. Método SCAD

En los últimos años se han desarrollado algunas generalizaciones y extensiones de las técnicas presentadas anteriormente, especialmente diseñadas para ciertas situaciones donde el método Lasso y la penalización L_q podrían no ser satisfactorias.

Las técnicas de penalización L_q con $0 \leq q < 1$, no satisfacen la condición de continuidad, de esparsidad y de insesgadez simultáneamente, por tanto como alternativa a los métodos anteriores se propone la penalización SCAD que se define de la siguiente manera,

$$p_\lambda^{SCAD}(\beta_j) = \begin{cases} \lambda|\beta_j| & \text{si } 0 \leq |\beta_j| \leq \lambda \\ -(\beta_j^2 - 2a\lambda|\beta_j| + \lambda^2)/(2(a-1)) & \text{si } \lambda \leq |\beta_j| \leq a\lambda \\ (a+1)\lambda^2/2 & \text{si } |\beta_j| \geq a\lambda \end{cases}$$

donde $a > 2$ y $\lambda > 0$.

El estimador SCAD se define entonces como el que minimiza la siguiente expresión

$$\sum_{i=1}^{n}(y_i - \boldsymbol{x_i'}\boldsymbol{\beta})^2 + \lambda \sum_{j=1}^{p} p_\lambda^{SCAD}(\beta_j) \tag{2.6}$$

y fue introducido por Fan y Li (2001). Estos autores recomiendan utilizar $a = 3.7$.

2.3. Métodos para seleccionar el parámetro de penalización λ

Como puede observarse todas estas técnicas de mínimos cuadrados penalizados dependen de un parámetro de penalización λ, que controla la importancia dada a la penalización en el proceso de optimización. Cuanto mayor es λ mayor es la penalización en los coeficientes de regresión y más son contraídos éstos hacia cero. Nótese que si $\lambda = 0$ la estimación coincide con la de mínimos cuadrados ordinarios.

Algunos de los métodos para seleccionar el parámetro λ son los siguientes:

- **CV (Validación cruzada):** Uno de los métodos más utilizados para estimar el parámetro λ es el método *k-fold cross-validation.*

  Denotamos por $T = \{(y_i, \boldsymbol{x_i}), i = 1, \ldots, n\}$ al conjunto de todos los datos, y a los conjuntos de entrenamiento y pueba por $T - T^v$ y T^v respectivamente, para $v = 1, \ldots, k$. Cada T^v está formado por una proporción fija de elementos de T seleccionados al azar.

  Sea $\widehat{\boldsymbol{\beta}}_{\boldsymbol{n}}^{(\boldsymbol{v})}$ el estimador de $\boldsymbol{\beta}$ utilizando el conjunto de datos de entrenamiento $T - T^v$. El estadístico de validación cruzada se define como

$$CV(\lambda) = \sum_{v=1}^{k} \sum_{(y_k, \boldsymbol{x_k}) \in T^v} \{y_k - \boldsymbol{x_k'} \widehat{\boldsymbol{\beta}}_{\boldsymbol{n}}^{(\boldsymbol{v})}(\lambda)\}^2.$$

  Por tanto, el estimador del parámetro λ se denotará como $\widehat{\lambda}$ y será el que minimiza la expresión $CV(\lambda)$.

- **GCV (Validación cruzada generalizada):** Otro de los métodos para seleccionar el parámetro λ es el de validación cruzada generalizada. El estadístico de dicho método se define como

$$GCV(\lambda) = \frac{1||\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}(\lambda)||^2}{n(1 - \frac{df(\lambda)}{n})^2},$$

  donde $df(\lambda) = tr(\boldsymbol{B}(\lambda))$, siendo $\boldsymbol{B}(\lambda)$ la matriz tal que $\widehat{Y} = \boldsymbol{B}(\lambda)\boldsymbol{Y}$. Así, el parámetro λ se estimará como

$$\widehat{\lambda} = arg\, \min_{\lambda}(GCV(\lambda)),$$

- **BIC**: Este criterio se define de la siguiente manera

$$BIC = 2L + \log(n) df(\lambda)$$

  donde L es la verosimilitud, n el tamaño de muestra y $df(\lambda)$ ya definido anteriormente. Esta medida tiene en cuenta el ajuste y a la vez compensa el exceso de parámetros.

Se trata de buscar un modelo cuyo BIC sea pequeño, pues en ese caso habría una verosimilitud grande y pocos parámetros.

- **AIC**: Este otro criterio se define como sigue:

$$AIC = 2L + 2df(\lambda)$$

donde L es la verosimilitud, n el tamaño de la muestra y $df(\lambda)$ ya definido anteriormente. Al igual que el criterio BIC esta medida tiene en cuenta el ajuste y a la vez compensa el exceso de parámetros. De nuevo se trata de buscar un modelo cuyo AIC sea pequeño, pues en ese caso habría una verosimilitud grande y pocos parámetros.

Capítulo 3

Comparativa de métodos de selección de variables: Teoría

3.1. Introducción

En el capítulo anterior ya hemos explicado la importancia que tiene la selección de variables en un modelo de regresión. Este capítulo estará dedicado al estudio de tres métodos para la selección de variables utilizando mínimos cuadrados penalizados, como son el método Lasso, el método Bridge y el método SCAD. Cada sección estudiará cada método por separado diferenciando en cada caso, si el número de variables p es fijo, o si en caso contrario el número de variables p_n crece hacia infinito cuando n tiende a infinito. Parece claro que hay algunas propiedades deseables que todo criterio de selección debiera satisfacer; esto es, existen algunas formas objetivas de comparar los distintos criterios de selección y concluir cual de ellos es el mejor, por tanto, en último lugar se pasará a un estudio comparativo del comportamiento de los distintos criterios de selección en base a las propiedades que cumplen cada uno de ellos.

En la siguiente sección se presentarán determinadas propiedades deseables que todo criterio de selección de variables debiera satisfacer.

3.2. Propiedades deseables

Partimos del modelo $y_i = x_{i1}\beta_1 + \ldots + x_{ip}\beta_p + \varepsilon_i$ para $i = 1, \ldots, n$, que escribiremos también como $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$. Existen propiedades deseables que todo criterio de selección de variables debiera satisfacer. A continuación se expondrán algunas de ellas, las dos primeras solo tienen sentido si el número de variables p es constante.

1. **Consistencia del estimador de los parámetros ($\widehat{\boldsymbol{\beta}}_{\boldsymbol{n}}$):**

   Se dice que el estimador $\widehat{\boldsymbol{\beta}}_{\boldsymbol{n}}$ es consistente cuando éste converge a su valor verdadero ($\boldsymbol{\beta}$) cuando el número de datos de la muestra tiende a infinito, es decir,

   $$\widehat{\boldsymbol{\beta}}_{\boldsymbol{n}} \xrightarrow{p} \boldsymbol{\beta} \text{ cuando } n \to \infty$$

2. **Normalidad asintótica:** Se tiene la normalidad asintótica de los estimadores de los coeficientes no nulos cuando

   $$\sqrt{n}(\widehat{\boldsymbol{\beta}}_{\mathbf{1}\boldsymbol{n}} - \boldsymbol{\beta}_{\mathbf{1}}) \to N(\mathbf{0}, \boldsymbol{V})$$

   siendo $\boldsymbol{V}$ una matriz de dimensión $k_n \times k_n$, donde k_n era el número de parámetros no nulos.

3. **Consistencia en la selección del modelo.**

   Se tiene la consistencia en la selección del modelo cuando

   $$P(\widehat{\boldsymbol{S}} = \boldsymbol{S}) \to 1,$$

   es decir, con probabilidad tendiendo a 1 se seleccionan bien los parámetros distintos de cero.

4. **Esparsidad**

   Una propiedad importante que cumplen algunos estimadores es que los coeficientes que son cero se fijan a cero con probabilidad tendiendo a 1, es decir,

$$P(\widehat{\boldsymbol{\beta}}_{\boldsymbol{2n}} = \mathbf{0}_{\boldsymbol{m_n}}) \to 1.$$

   Nótese que esta propiedad es más débil que la consistencia en la selección del modelo.

Si un estimador verifica las propiedades 2) y 3) se dice que dicho estimador cumple la **propiedad del oráculo**: esto es que el estimador de mínimos cuadrados penalizados funciona tan bien como si el submodelo correcto se conociera de antemano. Aunque habitualmente se dice que un estimador cumple la propiedad del oráculo si satisface las propiedades 2) y 3), algunos artículos también proponen que un estimador verifica la propiedad del oráculo cuando satisface las propiedades 2) y 4).

También sería deseable conocer la velocidad de convergencia del estimador de los parámetros:

$$||\widehat{\boldsymbol{\beta}}_{\boldsymbol{n}} - \boldsymbol{\beta}|| = O_p(n^{-\delta}).$$

Lo ideal es que $\delta = \frac{1}{2}$.

Trataremos los casos en que el número de variables p es fijo y el caso en el que el número de variables p_n crece a medida que el tamaño muestral n crece. Como ya se ha dicho anteriormente las propiedades 1) y 2) solo tienen sentido si el número de variables p es fijo, en caso de que el número de variables no sea fijo en lugar de 1) y 2) consideraremos las siguientes propiedades:

- 1'. **Consistencia del estimador de los parámetros ($\widehat{\boldsymbol{\beta}}_{\boldsymbol{n}}$):**

  Se dice que el estimador $\widehat{\boldsymbol{\beta}}_{\boldsymbol{n}}$ es consistente cuando éste converge en norma L_2 a su valor verdadero ($\boldsymbol{\beta}$) cuando el número de datos de la muestra tiende a infinito, es decir,

$$||\widehat{\boldsymbol{\beta}}_{\boldsymbol{n}} - \boldsymbol{\beta}|| \xrightarrow{p} 0 \text{ cuando } n \to \infty$$

donde $||.||$ es la norma L_2, esto es:

$$||(a_1, \ldots, a_p)|| = \sqrt{a_1^2 + \ldots + a_p^2}$$

- 2'. **Normalidad asintótica**: Se tiene la normalidad asintótica de los estimadores de los coeficientes no nulos cuando

$$\sqrt{n}\boldsymbol{A_n}(\widehat{\boldsymbol{\beta}}_{\mathbf{1}\boldsymbol{n}} - \boldsymbol{\beta_1}) \to N(\mathbf{0}, \boldsymbol{V})$$

donde $\boldsymbol{A_n}$ es una matriz $d \times k_n$ (k_n es el número de parámetros distintos de cero y $d \geq 1$) y $\boldsymbol{V}$ una matriz de dimensión $d \times d$.

Seguidamente citaremos las condiciones más usuales que se exigirán para que los estimadores cumplan determinadas propiedades.

3.3. Condiciones clave

Como hemos dicho antes, en esta sección citaremos algunas condiciones clave que se exigirán para que los estimadores de los distintos métodos de regresión penalizada, cumplan determinadas propiedades. No todas estas condiciones deben ser satisfechas de manera simultánea, sino que cada método requerirá la verificación de algunas de ellas. Si bien se necesitan más condiciones, no serán mostradas aquí debido a su carácter eminentemente técnico.

1. $(y_i, \boldsymbol{x}_i)$ son i.i.d, es decir, independientes e igualmente distribuidas.
2. $\varepsilon_i^{'}$ tienen media 0 y varianza σ^2.

   Si el número de variables p es fijo:

3. $C_n = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x_i}\boldsymbol{x_i'} \to \boldsymbol{C}$ siendo $\boldsymbol{C}$ una matriz $p \times p$ no singular.

   Si el número de variables p_n crece con el tamaño de muestra n:

4. Denotamos por $\boldsymbol{x}_i = (\boldsymbol{w}_i, \boldsymbol{z}_i)$ donde $\boldsymbol{w}_i$ contiene las k_n covariables correspondientes a los coeficientes distintos de cero, y $\boldsymbol{z}_i$ contiene las m_n variables correspondientes a los coeficientes cero.

   Sea $\boldsymbol{X}_n$, $\boldsymbol{X}_{1n}$ y $\boldsymbol{X}_{2n}$ matrices cuyas transpuestas son $\boldsymbol{X}_n^{'} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$, $\boldsymbol{X}_{1n}^{'} = (\boldsymbol{w}_1, \ldots, \boldsymbol{w}_n)$ y $\boldsymbol{X}_{2n}^{'} = (\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n)$. Por otro lado, denotamos por $\boldsymbol{\Sigma_n} = n^{-1}\boldsymbol{X}_n^{'}\boldsymbol{X}_n$, y $\boldsymbol{\Sigma_{1n}} = n^{-1}\boldsymbol{X}_{1n}^{'}\boldsymbol{X}_{1n}$,

   Entonces,

$$\rho_{1n} > 0 \text{ para todo } n$$

   siendo ρ_{1n} el autovalor más pequeño de $\boldsymbol{\Sigma_n}$.

5. Existen constantes $0 < \tau_1 < \tau_2 < \infty$ tales que $\tau_1 \leq \tau_{1n} \leq \tau_{2n} \leq \tau_2$ para todo n, siendo τ_{1n} y τ_{2n} el autovalor más pequeño y más grande de $\boldsymbol{\Sigma_{1n}}$, respectivamente.

6. Existen constantes $0 < b_0 < b_1 < \infty$ tal que

$$b_0 \leq \text{mín}\{|\beta_{1j}|,\ 1 \leq j \leq k_n\} \leq \text{máx}\{|\beta_{1j}|,\ 1 \leq j \leq k_n\} \leq b_1$$

7. $$\frac{p_n}{n} \to 0 \quad \text{cuando } n \to \infty$$

8. $$\frac{p_n^2}{n} \to 0 \quad \text{cuando } n \to \infty$$

9. $$\frac{\log(p_n)}{n} \to\ 0 \quad \text{cuando } n \to \infty$$

3.4. Estudio teórico del método Lasso

Esta sección estará dedicada al estudio del método Lasso, tanto si el número de variables p es fijo como si el número de variables p_n crece cuando el tamaño muestral n crece.

Consideremos el modelo de regresión lineal

$$y_i = x_{i1}\beta_1 + \ldots + x_{ip}\beta_p + \varepsilon_i.$$

El método Lasso consistirá en estimar el vector de parámetros $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$ minimizando el siguiente criterio de mínimos cuadrados penalizados

$$\sum_{i=1}^{n}(y_i - \boldsymbol{x_i'}\boldsymbol{\beta})^2 + \lambda \sum_{j=1}^{p} |\beta_j| \tag{3.1}$$

para un λ dado.

Denotaremos por $\widehat{\boldsymbol{\beta}}_{\boldsymbol{n}}$ al estimador de los parámetros ($\boldsymbol{\beta}$), es decir el vector donde se alcanza el mínimo de la expresión (3.1).

3.4.1. Número de variables p fijo

En esta sección abordaremos el estudio de determinadas propiedades que bajo ciertas condiciones cumple el método Lasso en el caso de que el número de variables p sea fijo.

Propiedades del método Lasso

Comenzaremos citando las propiedades más importantes de dicho estimador como son la consistencia del estimador de los parámetros y la normalidad asintótica de los estimadores de los coeficientes no nulos.

El siguiente resultado muestra que el estimador de parámetros $\widehat{\boldsymbol{\beta}}_{\boldsymbol{n}}$ es consistente.

Teorema 3.4.1. *(KNIGHT AND FU 2000)*

Consistencia del estimador de los parámetros ($\widehat{\boldsymbol{\beta}}_{\boldsymbol{n}}$)*: Bajo determinadas condiciones (entre ellas la condiciones clave 1, 2 y 3 de la Sección 3.3) se verifica que*

$$\widehat{\boldsymbol{\beta}}_{\boldsymbol{n}} \xrightarrow{p} \boldsymbol{\beta}$$

y por tanto, se tiene la consistencia del estimador de los parámetros $\widehat{\boldsymbol{\beta}}_{\boldsymbol{n}}$.

A continuación, se muestra la normalidad asintótica de los estimadores de los coeficientes no nulos.

Teorema 3.4.2. *(KNIGHT AND FU 2000)*

***Normalidad asintótica**: Nuevamente bajo ciertas condiciones (entre ellas la condiciones clave 1, 2 y 3 de la Sección 3.3) se tiene que*

$$\sqrt{n}(\widehat{\boldsymbol{\beta}}_{\mathbf{1}\boldsymbol{n}} - \boldsymbol{\beta}_1) \to N(\mathbf{0}, \boldsymbol{P}\boldsymbol{V}\boldsymbol{P}')$$

donde $\boldsymbol{P}$ es una matriz de dimesión $k_n \times p$ siendo la fila i-ésima de $\boldsymbol{P}$, $P_i = (0, \ldots, 0, \underbrace{1}_{i}, 0 \ldots, 0)$ para $i = 1, \ldots, k_n$, es decir, un vector de ceros, salvo la posición i-ésima que vale 1.

Por otra parte, $\boldsymbol{V} = \sigma^2 \boldsymbol{C}^{-1}$ (la definición de $\boldsymbol{C}$ puede verse en la condición clave 3 de la Sección 3.3).

Teorema 3.4.3. *(KNIGHT AND FU 2000)*

***Razón de convergencia**: Bajo ciertas condiciones, (entre ellas la condiciones clave 1, 2 y 3 de la Sección 3.3) se tiene que*

$$||\widehat{\boldsymbol{\beta}}_{\boldsymbol{n}} - \boldsymbol{\beta}|| = O_p(n^{-\frac{1}{2}})$$

Seguidamente estudiaremos el caso en que el número de variables $p = p_n$ aumenta con el tamaño de muestra n.

3.4.2. Número de variables $p = p_n$ tendiendo a infinito cuando n tiende a infinito ($p_n < n$)

En esta sección estudiaremos determinadas propiedades que bajo ciertas condiciones cumple el estimador Lasso cuando el número de variables $p = p_n$ crece con el tamaño de muestra n, pero se mantiene por debajo de n.

Propiedades del método Lasso

La propiedad que cumplirá el método Lasso en este caso es la consistencia del estimador de los parámetros que se muestra en el siguiente resultado:

Teorema 3.4.4. *(HUANG, HOROWITZ Y MA 2008)*

Consistencia del estimador de los parámetros $(\widehat{\boldsymbol{\beta}}_{\boldsymbol{n}})$*: Bajo determinadas condiciones (entre ellas las condiciones clave 1, 2, 4, 6 y 7 de la Sección 3.3) se verifica que*

$$||\widehat{\boldsymbol{\beta}}_{\boldsymbol{n}} - \boldsymbol{\beta}|| \xrightarrow{p} 0,$$

es decir, el estimador de los parámetros $\widehat{\boldsymbol{\beta}}_{\boldsymbol{n}}$ *es consistente.*

Teorema 3.4.5. *(HUANG, HOROWITZ Y MA 2008)*

Razón de convergencia*: Es deseable conocer la velocidad de convergencia del estimador de los parámetros, que en este caso, bajo determinadas condiciones (entre ellas las condiciones clave 1 y 2 de la Sección 3.3), es la siguiente*

$$||\widehat{\boldsymbol{\beta}}_{\boldsymbol{n}} - \boldsymbol{\beta}|| = O_p\left(\sqrt{\frac{p_n}{n}}\right)$$

3.4.3. Consistencia en la selección del modelo y condición de irrepresentabilidad

El método Lasso no es consistente en la selección del modelo. Para alcanzar dicha consistencia tanto cuando el número de variables p es fijo como cuando aumenta a medida que aumenta el tamaño de muestra n, es necesario imponer una condición muy fuerte, conocida como la Condición de Irrepresentabilidad.

La condición de irrepresentabilidad que depende principalmente de la covarianza de las variables predictoras, sostiene que el método Lasso selecciona el verdadero modelo consistentemente si y solo si las variables predictoras que no están en el verdadero modelo son "irrepresentables" por predictores que están en el verdadero modelo.

Sin pérdida de generalidad, asumimos que $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_{k_n}, \beta_{k_n+1}, \ldots, \beta_p)'$ donde $\beta_j \neq 0$ para $j = 1, \ldots, k_n$ y $\beta_j = 0$ para $j = k_n+1, \ldots, p$. Sean $\boldsymbol{X_1}$ y $\boldsymbol{X_2}$ las matrices formadas por

las primeras k_n y las $m_n = p - k_n$ columnas de $\boldsymbol{X}$ respectivamente, y $\boldsymbol{C} = \frac{1}{n}\boldsymbol{X}'\boldsymbol{X}$. Usaremos la notación $\boldsymbol{C_{11}} = \frac{1}{n}\boldsymbol{X_1}'\boldsymbol{X_1}, \boldsymbol{C_{22}} = \frac{1}{n}\boldsymbol{X_2}'\boldsymbol{X_2},\ \boldsymbol{C_{12}} = \frac{1}{n}\boldsymbol{X_1}'\boldsymbol{X_2}$ y $\boldsymbol{C_{21}} = \frac{1}{n}\boldsymbol{X_2}'\boldsymbol{X_1}$.

De esta manera $\boldsymbol{C}$ puede expresarse como

$$\boldsymbol{C} = \begin{pmatrix} \boldsymbol{C_{11}} & \boldsymbol{C_{12}} \\ \boldsymbol{C_{21}} & \boldsymbol{C_{22}} \end{pmatrix}.$$

Asumiendo que $\boldsymbol{C_{11}}$ es invertible, definimos la **Condición de Irrepresentabilidad** como sigue:

Existe un vector constante positivo η tal que

$$|\boldsymbol{C_{21}}(\boldsymbol{C_{11}})^{-1}\text{sign}(\boldsymbol{\beta_1})| \leq \mathbf{1} - \eta.$$

Una vez definida esta condición presentaremos el siguiente resultado que nos da la consistencia en la selección del modelo del método Lasso, tanto si el número de variables p es fijo como si aumenta al aumentar el tamaño de muestra.

Teorema 3.4.6. *(ZHAO Y YU 2006)*

***Consistencia en la selección del modelo**: Bajo ciertas condiciones (entre ellas la Condición de Irrepresentabilidad y las condiciones claves 1, 2 y 7 de la Sección 3.3, ésta última solo en el caso de que p no sea fijo) se tiene que el método Lasso es consistente en la selección del modelo, es decir se cumple que*

$$P(\widehat{\boldsymbol{S}} = \boldsymbol{S}) \to 1,$$

luego con probabilidad convergiendo a 1 se seleccionan bien los parámetros no nulos.

3.4.4. Método Lasso Adaptado

Una manera de obtener la consistencia en la selección del modelo sin necesidad de imponer fuertes condiciones es adaptar o modificar el método Lasso.

Por tanto vamos a proponer una nueva versión del método Lasso, llamada método

Lasso Adaptado, que utilizará pesos adaptados para penalizar diferentes coeficientes en la penalización L_1.

Consideremos el modelo de regresión lineal

$$y_i = x_{i1}\beta_1 + \ldots + x_{ip}\beta_p + \varepsilon_i.$$

El método Lasso Adaptado consistirá en estimar el vector de parámetros $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$ minimizando el siguiente criterio de mínimos cuadrados penalizados

$$\sum_{i=1}^{n}(y_i - \boldsymbol{x_i'}\boldsymbol{\beta})^2 + \lambda\sum_{j=1}^{p} w_j|\beta_j|$$

donde $\boldsymbol{w} = (w_1, \ldots, w_p)'$ es un vector de pesos conocido.

Definimos el vector de pesos como $\widehat{w} = \frac{1}{|\widehat{\boldsymbol{\beta}}|^\gamma}$, entonces la estimación $\widehat{\boldsymbol{\beta}}_{\boldsymbol{n}}$ de $\boldsymbol{\beta}$ viene dada por la que minimiza la siguiente expresión

$$\sum_{i=1}^{n}(y_i - \boldsymbol{x_i'}\boldsymbol{\beta})^2 + \lambda\sum_{j=1}^{p} \widehat{w_j}|\beta_j|$$

Propiedades del método Lasso Adaptado

En esta sección mostraremos que eligiendo adecuadamente el parámetro λ, para el método Lasso Adaptado se tiene la consistencia del estimador de los parámetros, la normalidad asintótica de los coeficientes no nulos y la consistencia en la selección del modelo cuando el número de variables p es fijo.

Teorema 3.4.7. *(ZOU 2006)*

***Consistencia del estimador de los parámetros**: Bajo determinadas condiciones (entre ellas las condiciones clave 1, 2 y 3 de la Sección 3.3) se verifica que*

$$||\widehat{\boldsymbol{\beta}}_{\boldsymbol{n}} - \boldsymbol{\beta}|| \xrightarrow{p} 0,$$

es decir, el estimador de los parámetros $\widehat{\boldsymbol{\beta}}_{\boldsymbol{n}}$ es consistente.

Teorema 3.4.8. *(ZOU 2006)*

***Normalidad asintótica**: Bajo determinadas condiciones (entre ellas las condiciones clave 1, 2 y 3 de la Sección 3.3) se tiene la normalidad asintótica de los estimadores de los coeficientes no nulos*

$$\sqrt{n}(\widehat{\boldsymbol{\beta}}_{\mathbf{1}\boldsymbol{n}} - \boldsymbol{\beta}_{\mathbf{1}}) \to N(\mathbf{0}, \boldsymbol{V})$$

donde $\boldsymbol{V}$ es una matriz de dimensión $k_n \times k_n$, donde k_n era el número de parámetros no nulos.

Teorema 3.4.9. *(ZOU 2006)*

Consistencia en la selección del modelo

Bajo determinadas condiciones (entre ellas las condiciones clave 1, 2 y 3) se tiene la consistencia en la selección del modelo, es decir, se cumple

$$P(\widehat{\boldsymbol{S}} = \boldsymbol{S}) \to 1,$$

luego con probabilidad 1 se seleccionan bien los parámetros distintos de 0.

Como podemos observar para el método Lasso Adaptado se tiene la normalidad asintótica de los estimadores de los coeficientes no nulos y la consistencia en la selección del modelo, luego cumple la propiedad del oráculo, es decir, funciona tan bien como si los coeficientes nulos se conocieran de antemano.

A continuación ilustraremos gráficamente el hecho de por que el método Lasso a menudo produce estimaciones de los parámetros nulas.

Geometría del método Lasso

En esta sección daremos una idea gráfica de por que el método Lasso a menudo produce estimaciones de β_k nulas (y por tanto selecciona variables) y en cambio el método ridge que usa la contracción $\sum \beta_j^2 \leq t$ en lugar de $\sum |\beta| \leq t$, no.

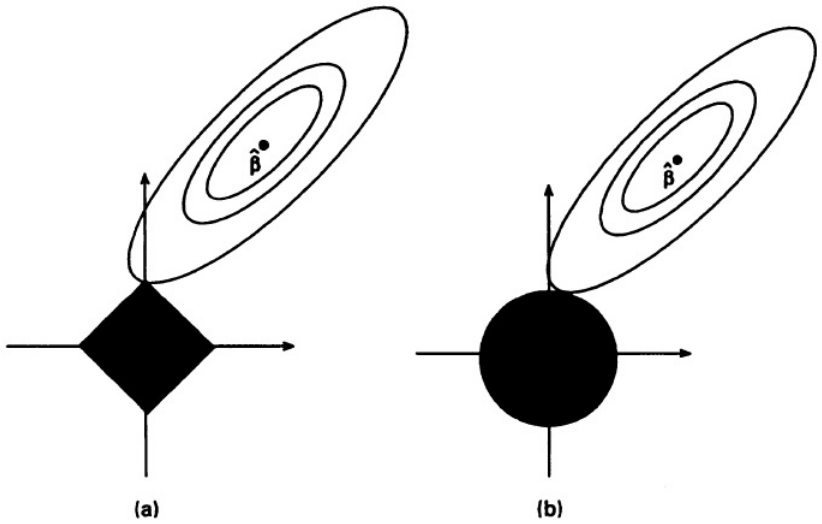

Figura 3.1: Representación gráfica de la estimación de los parámetros β_1 y β_2 para el método Lasso (a) y el método Ridge (b).

El criterio $\sum_{i=1}^{n}(y_i - \boldsymbol{x_i'}\boldsymbol{\beta})^2$ es igual a la función cuadrática

$$(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}_{\boldsymbol{n}}^{\mathbf{0}})' \boldsymbol{X'X}(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}_{\boldsymbol{n}}^{\mathbf{0}}).$$

Los contornos elípticos de esta función se muestran mediante las curvas completas en la Figura 3.1 (a), que se centran en las estimaciones OLS. La región de restricción para el método Lasso es el cuadrado rotado. La solución Lasso es el primer punto donde los contornos tocan al cuadrado, y esto a veces ocurre en una esquina, que corresponde a un coeficiente nulo. El gráfico para la regresión ridge se muestra en la Figura 3.1 (b), no hay esquinas para que los contornos toquen y por tanto, raramente las soluciones resultan nulas.

3.5. Estudio teórico del método bridge

En esta sección estudiaremos el comportamiento de los estimadores de regresión que minimizan la suma de cuadrados residual más una penalización proporcional a $\sum |\beta_j|^\gamma$ para algún $\gamma > 0$. Estos estimadores incluyen al estimador Lasso, que es un caso especial

cuando $\gamma = 1$. Estudiaremos el caso en que el número de variables p sea fijo y el caso en que el número de variables p_n aumente a medida que aumenta el tamaño muestral n.

Consideremos el modelo de regresión lineal.

$$y_i = x_{i1}\beta_1 + \ldots + x_{ip}\beta_p + \varepsilon_i.$$

El método Bridge consistirá en estimar el vector de parámetros $\boldsymbol{\beta}$ minimizando el siguiente criterio de mínimos cuadrados penalizados

$$\sum_{i=1}^{n}(y_i - \boldsymbol{x_i'}\boldsymbol{\beta})^2 + \lambda \sum_{j=1}^{p} |\beta_j|^\gamma \tag{3.2}$$

para un λ dado, donde $\gamma > 0$.

Tales estimadores fueron llamados **estimadores Bridge** por Frank y Friedman (1993), quienes los introdujeron como una generalización de la regresión ridge (cuando $\gamma = 2$). Nótese que cuando $\gamma = 1$ tenemos el estimador Lasso.

Como habitualmente, denotaremos al estimador del vector de parámetros $\boldsymbol{\beta}$ que minimiza la expresión (3.2) por $\widehat{\boldsymbol{\beta}}_{\boldsymbol{n}}$.

Seguidamente estudiaremos el caso en que el número de variables p es fijo.

3.5.1. Número de variables p fijo

Como hemos dicho en esta sección abordaremos el estudio de las propiedades más importantes que bajo ciertas condiciones cumple el estimador Bridge en el caso en el que el número de variables p es fijo.

Propiedades de los estimadores Bridge

Las propiedades más relevantes son la consistencia del estimador de los parámetros, la normalidad asintótica de los estimadores de los coeficientes no nulos y la propiedad de identificar los coeficientes nulos con probabilidad tendiendo a 1 (ésto solo si $0 < \gamma < 1$).

El siguiente resultado muestra que el estimador de los parámetros $\widehat{\boldsymbol{\beta}}_{\boldsymbol{n}}$ es consistente.

Teorema 3.5.1. *(KNIGHT AND FU 2000)*

Consistencia del estimador de los parámetros ($\widehat{\boldsymbol{\beta}}_{\boldsymbol{n}}$)*: Bajo ciertas condiciones (entre ellas las condiciones clave 1, 2 y 3 de la Sección 3.3) se cumple que*

$$\widehat{\boldsymbol{\beta}}_{\boldsymbol{n}} \xrightarrow{p} \boldsymbol{\beta}$$

y por tanto, el estimador de los parámetros $\widehat{\boldsymbol{\beta}}_{\boldsymbol{n}}$ *es consistente.*

El siguiente resultado muestra que existe normalidad asintótica de los estimadores de los coeficientes no nulos cuando $\gamma \geq 1$.

Teorema 3.5.2. *(KNIGHT AND FU 2000)*

Normalidad asintótica*: Suponemos que* $\gamma \geq 1$*. Bajo ciertas condiciones (entre ellas las condiciones clave 1, 2 y 3 de la Sección 3.3) se tiene que*

$$\sqrt{n}(\widehat{\boldsymbol{\beta}}_{\mathbf{1}\boldsymbol{n}} - \boldsymbol{\beta}_{\mathbf{1}}) \to N(\mathbf{0}, \boldsymbol{P}\boldsymbol{V}\boldsymbol{P}^{'})$$

donde $\boldsymbol{P}$ *es una matriz de dimesión* $k_n \times p$ *siendo la fila i-ésima de* $\boldsymbol{P}$*,* $P_i = (0, \ldots, 0, \underbrace{1}_{i}, 0 \ldots, 0)$ *para* $i = 1, \ldots, k_n$*, es decir, un vector de ceros, salvo la posición i-ésima que vale 1.*

Por otra parte, $\boldsymbol{V} = \sigma^2 \boldsymbol{C}^{-1}$ *(la definición de la matriz* $\boldsymbol{C}$ *puede verse en la condición clave 3 de la Sección 3.3).*

Teorema 3.5.3. *(HUANG, HOROWITZ Y MA 2008)*

Normalidad asintótica: *Suponemos que* $0 < \gamma < 1$*. Entonces bajo ciertas condiciones (entre ellas las condiciones clave 1, 2 y 3 de la Sección 3.3) se tiene:*

$$\sqrt{n}(\widehat{\boldsymbol{\beta}}_{\mathbf{1}\boldsymbol{n}} - \boldsymbol{\beta}_{\mathbf{1}}) \to N(0, \Sigma)$$

donde Σ *es la matriz de varianzas y covarianzas de la distribución normal.*

Teorema 3.5.4. *(HUANG, HOROWITZ Y MA 2008)*

Supongamos que $0 < \gamma < 1$ *entonces bajo ciertas condiciones (entre ellas las condiciones clave 1, 2 y 3 de la Sección 3.3) se cumple lo siguiente:*

$$P(\widehat{\boldsymbol{\beta}}_{\boldsymbol{2n}} = \mathbf{0}_{\boldsymbol{m_n}}) \to 1$$

es decir con probabilidad convergiendo a 1 se estiman correctamente los coeficientes nulos.

Nótese que cuando el parámetro $0 < \gamma < 1$ y el número de variables p es fijo, se tiene la normalidad asintótica de los estimadores de los coeficientes no nulos y también la propiedad de identificar correctamente los coeficientes nulos con probabilidad tendiendo a 1. Por tanto, si $0 < \gamma < 1$ el estimador Bridge cumple la propiedad del oráculo, es decir, funciona tan bien como si los coeficientes nulos fueran conocidos de antemano.

Teorema 3.5.5. *(HUANG, HOROWITZ Y MA 2008)*

***Razón de convergencia**: Bajo ciertas condiciones (entre ellas las condiciones clave 1, 2 y 3 de la Sección 3.3), se tiene que*

$$||\widehat{\boldsymbol{\beta}}_{\boldsymbol{n}} - \boldsymbol{\beta}|| = O_p(n^{-\frac{1}{2}})$$

Seguidamente se estudiarán las propiedades del estimador Bridge cuando el número de variables $p = p_n$ aumenta con el tamaño de muestra n.

3.5.2. Número de variables $p = p_n$ tendiendo a infinito cuando n tiende a infinito $(p_n < n)$

En esta sección estudiaremos determinadas propiedades que bajo ciertas condiciones verifica el estimador Bridge cuando el número de variables $p = p_n$ crece con el tamaño de muestra n, pero se mantiene por debajo de n.

Propiedades de los estimadores Bridge

Algunas de las propiedades más relevantes son la consistencia del estimador de los parámetros, y cuando $0 < \gamma < 1$ la normalidad asintótica de los estimadores de los coeficientes no nulos y la propiedad de identificar correctamente los coeficientes nulos con probabilidad tendiendo a 1.

El siguiente Teorema muestra que si $\gamma > 0$ el estimador de los parámetros $\widehat{\boldsymbol{\beta}}_{\boldsymbol{n}}$ es consistente.

Teorema 3.5.6. *(HUANG, HOROWITZ AND MA 2008)*

Consistencia del estimador de los parámetros ($\widehat{\boldsymbol{\beta}}_{\boldsymbol{n}}$)*: Suponiendo que* $\gamma > 0$*, bajo ciertas condiciones, (entre ellas las condiciones clave 1, 2, 4, 6 y 7 de la Sección 3.3) se tiene que*

$$||\widehat{\boldsymbol{\beta}}_{\boldsymbol{n}} - \boldsymbol{\beta}|| \xrightarrow{p} 0,$$

es decir, el estimador de los parámetros $\widehat{\boldsymbol{\beta}}_{\boldsymbol{n}}$ *es consistente.*

Teorema 3.5.7. *(HUANG, HOROWITZ AND MA 2008)*

Normalidad asintótica: *Supongamos que* $0 < \gamma < 1$ *entonces bajo ciertas condiciones (entre ellas las condiciones clave 1, 2, 4, 5, 6 y 7 de la Seción 3.3) se cumple lo siguiente:*

$$\sqrt{n}\boldsymbol{A}_{\boldsymbol{n}}(\widehat{\boldsymbol{\beta}}_{\mathbf{1}\boldsymbol{n}} - \boldsymbol{\beta}_{\mathbf{1}}) \to N(0, 1)$$

donde $\boldsymbol{A}_{\boldsymbol{n}}$ *es una matriz de dimensión* $d \times k_n$ *con* $d = 1$ *y siendo* k_n *el número de coeficientes distintos de cero.*

Teorema 3.5.8. *(HUANG, HOROWITZ AND MA 2008):*

Supongamos que $0 < \gamma < 1$ *entonces bajo ciertas condiciones (entre ellas las condiciones clave 1, 2, 4, 5, 6 y 7 de la Seción 3.3) se tiene lo siguiente*

$$P(\widehat{\boldsymbol{\beta}}_{\mathbf{2}\boldsymbol{n}} = \mathbf{0}_{\boldsymbol{m_n}}) \to 1$$

es decir con probabilidad 1 se estiman correctamente los coeficientes nulos.

Como hemos visto, si el número de variables p crece a medida que aumenta n y $0 < \gamma < 1$, se tiene la normalidad asintótica de los coeficientes no nulos y la propiedad de identificar correctamente los coeficientes nulos con probabilidad tendiendo a 1. Luego se dice que el estimador Bridge cumple la propiedad del oráculo.

Teorema 3.5.9. *(HUANG, HOROWITZ Y MA 2008)*

***Razón de convergencia**: Es deseable conocer la velocidad de convergencia del estimador, que en este caso bajo determinadas condiciones (entre ellas las condiciones clave 1 y 2 de la Sección 3.3) es la siguiente*

$$||\widehat{\boldsymbol{\beta}}_{\boldsymbol{n}} - \boldsymbol{\beta}|| = O_p\left(\sqrt{\frac{p_n}{n}}\right)$$

3.5.3. Número de variables $p = p_n$ tendiendo a infinito cuando n tiende a infinito ($p_n > n$)

En la sección anterior hemos visto que el estimador Bridge cuando $0 < \gamma < 1$ puede seleccionar correctamente covariables con coeficientes nulos solo si $p_n < n$. Por tanto, en esta sección estudiaremos una versión modificada del estimador Bridge, conocida como el estimador Bridge Marginal para contemplar el caso de que el número de covariables sea mayor que el tamaño de muestra ($p_n > n$). En este caso mostraremos que bajo una condición de ortogonalidad parcial según la cual las covariables de los coeficientes nulos están incorrelacionadas o débilmente correlacionadas con las covariables de los coeficientes no nulos, el estimador Bridge Marginal puede distinguir consistentemente entre coeficientes nulos y no nulos, es decir se conseguirá la consistencia en la selección del modelo.

Condición de Ortogonalidad Parcial: Existe una constante $c_0 > 0$ tal que

$$|n^{-1/2}\sum_{i=1}^{n} x_{ij}x_{ik}| \leq c_0 \quad j \in J_n, k \in K_n,$$

para todo n suficientemente grande, donde $K_n = \{1, \ldots, k_n\}$ siendo k_n el número de variables con coeficientes distintos de cero, y $J_n = \{k_n + 1, \ldots, p_n\}$, siendo p_n el número de

variables.

El método Bridge Marginal consistirá en estimar el vector de parámetros $\boldsymbol{\beta}$ minimizando la función objetivo

$$U_n(\beta) = \sum_{j=1}^{p}\sum_{i=1}^{n}(y_i - x_{ij}\beta_j)^2 + \lambda\sum_{j=1}^{p}|\beta_j|^{\gamma} \tag{3.3}$$

Propiedades asintóticas del estimador bridge marginal bajo la condición de ortogonalidad parcial

En esta sección se estudiarán las propiedades del estimador bridge marginal bajo la condición de ortogonalidad parcial.

Teorema 3.5.10. *(HUANG, HOROWITZ AND MA 2008)*

***Consistencia en la selección del modelo**: Bajo ciertas condiciones, (entre ellas las condiciones clave 1, 2 y 9 de la Sección 3.3 y la condición de ortogonalidad parcial), y suponiendo que* $0 < \gamma < 1$ *se tiene que*

$$P(\widehat{\boldsymbol{\beta}}_{\boldsymbol{2n}} = \mathbf{0}_{\boldsymbol{m_n}}) \to 1 \quad y\ P(\widehat{\beta}_k \neq 0, k \in K_n) \to 1$$

Este teorema nos dice que el estimador bridge marginal puede distinguir correctamente entre covariables con coeficientes nulos y covariables con coeficientes distintos de cero, es decir, es consistente en la selección del modelo. Sin embargo, los estimadores de los coeficientes distintos de cero no son consistentes. Para obtener estimadores consistentes, usaremos un método en dos pasos. Primero usaremos el estimador bridge marginal para seleccionar las covariables con coeficientes distintos de cero. Entonces estimamos el modelo de regresión con las covariables seleccionadas. En el segundo paso, cualquier método de regresión razonable puede ser usado. Usaremos el estimador bridge para estimar $\boldsymbol{\beta_1}$, el vector de coeficientes distintos de cero pues sabemos que dicho estimador es consistente en la estimación de los parámetros.

Como los coeficientes nulos son identificados correctamente con probabilidad convergiendo a 1, asumimos que solo las covariables con coeficientes distintos de cero son incluidas en

el modelo para el análisis asintótico de la estimación en el segundo paso.

Sea $\widehat{\boldsymbol{\beta}}^{*}_{\mathbf{1}\boldsymbol{n}}$ el estimador en este paso. Dicho estimador se puede definir como el valor que minimiza

$$U_n^*(\boldsymbol{\beta_1}) = \sum_{i=1}^{n}(y_i - \boldsymbol{w_i'}\boldsymbol{\beta_1})^2 + \lambda^* \sum_{j=1}^{k_n} |\beta_{1j}|^{\gamma} \tag{3.4}$$

donde $\boldsymbol{\beta_1} = (\beta_1, \dots \beta_{k_n})'$ y $\boldsymbol{w_i}$ es un vector de dimensión $k_n \times 1$ que contiene las k_n covariables con coeficientes distintos de cero.

Teorema 3.5.11. *(HUANG, HOROWITZ AND MA 2008)*

***Normalidad asintótica**: Bajo ciertas condiciones (entre ellas las condiciones clave 1, 2 y 9 de la Sección 3.3 y la condición de ortogonalidad parcial) y suponiendo que* $0 < \gamma < 1$*, se tiene que*

$$\sqrt{n}\boldsymbol{A_n}(\widehat{\boldsymbol{\beta}}^{\boldsymbol{*}\prime}_{\mathbf{1}\boldsymbol{n}} - \boldsymbol{\beta}^{\prime}_{\mathbf{1}}) \to N(0,1)$$

donde $\boldsymbol{A_n}$ *es una matriz de dimensión* $d \times k_n$ *con* $d = 1$ *y* k_n *el número de parámetros distintos de cero.*

A continuación estudiaremos otro de los métodos de regresión penalizada, el método SCAD. Lo estudiaremos tanto cuando el número de variables p es fijo como cuando el número de variables p_n aumenta al aumentar el tamaño de muestra n.

3.6. Estudio teórico del método SCAD

En los últimos años se han desarrollado algunas generalizaciones y extensiones de las técnicas presentadas anteriormente, especialmente diseñadas para ciertas situaciones donde el método Lasso y la penalización L_q podrían no ser satisfactorias.

Las técnicas de penalización L_q con $0 \leq q < 1$, no satisfacen simultáneamente las condiciones de insesgadez, esparsidad y continuidad. Como alternativa, se propone la penalización SCAD. En esta sección estudiaremos dicho método. Estudiaremos el caso en que el número de variables p es fijo, y el caso en que el número de variables p_n crece con el tamaño de muestra n.

Consideremos el modelo de regresión lineal

$$y_i = x_{i1}\beta_1 + \ldots + x_{ip}\beta_p + \varepsilon_i.$$

El estimador SCAD, se define entonces como el que minimiza la siguiente expresión

$$\sum_{i=1}^{n}(y_i - \boldsymbol{x_i'}\boldsymbol{\beta})^2 + \lambda \sum_{j=1}^{p} p_\lambda^{SCAD}(\beta_j) \tag{3.5}$$

para un λ dado, donde la penalización SCAD se define de la siguiente manera

$$p_\lambda^{SCAD}(\beta_j) = \begin{cases} \lambda|\beta_j| & \text{si } 0 \leq |\beta_j| \leq \lambda \\ -(\beta_j^2 - 2a\lambda|\beta_j| + \lambda^2)/(2(a-1)) & \text{si } \lambda \leq |\beta_j| \leq a\lambda \\ (a+1)\lambda^2/2 & \text{si } |\beta_j| \geq a\lambda \end{cases}$$

siendo $a > 2$ y λ parámetros de ajuste.

Denotaremos al mínimo de la expresión (3.5), es decir, al estimador de los parámetros como $\widehat{\boldsymbol{\beta}}_{\boldsymbol{n}}$.

3.6.1. Número de variables p fijo

En esta sección abordaremos el estudio de determinadas propiedades que bajo ciertas condiciones verifica el método SCAD cuando el número de variables p es fijo.

Propiedades del estimador SCAD

Las propiedades más destacadas de este estimador son la consistencia del estimador de los parámetros, la normalidad asintótica de los estimadores de los coeficientes no nulos y la propiedad de identificar correctamente los coeficientes nulos.

El siguiente resultado muestra que el estimador de los parámetros $\widehat{\boldsymbol{\beta}}_n$ es consistente.

Teorema 3.6.1. *(FAN AND LI 2001)*

Consistencia del estimador de los parámetros $(\widehat{\boldsymbol{\beta}}_{\boldsymbol{n}})$*: Bajo ciertas condiciones (entre ellas las condiciones clave 1, 2 y 3 de la Sección 3.3) se cumple que*

$$\widehat{\boldsymbol{\beta}}_n \xrightarrow{p} \boldsymbol{\beta}$$

y por tanto, el estimador de los parámetros $\widehat{\boldsymbol{\beta}}_n$ *es consistente.*

Teorema 3.6.2. *(FAN AND LI 2001)*

Normalidad asintótica*: Bajo ciertas condiciones (entre ellas las condiciones clave 1, 2 y 3 de la Sección 3.3) se tiene la normalidad asintótica de los estimadores de los coeficientes no nulos, es decir,*

$$\sqrt{n}(\widehat{\boldsymbol{\beta}}_{\mathbf{1}\boldsymbol{n}} - \boldsymbol{\beta}_{\mathbf{1}}) \to N(0, \boldsymbol{V})$$

donde $\boldsymbol{V}$ *es una matriz de tamaño* $k_n \times k_n$ *.*

Teorema 3.6.3. *(FAN AND LI 2001)*

Bajo ciertas condiciones (entre ellas las condiciones clave 1, 2 y 3 de la Sección 3.3) se cumple que

$$P(\widehat{\boldsymbol{\beta}}_{\mathbf{2}\boldsymbol{n}} = \mathbf{0}_{\boldsymbol{m_n}}) \to 1,$$

es decir, con probabilidad 1 se identifican correctamente los coeficientes nulos.

Como podemos observar si el número de variables p es fijo, para el estimador SCAD se tiene la normalidad asintótica de los coeficientes no nulos y la propiedad de identificar correctamente los coeficientes nulos. Por tanto se dice que el estimador SCAD cumple la propiedad del oráculo, es decir, funciona tan bien como si los coeficientes nulos se conocieran de antemano.

Teorema 3.6.4. *(FAN AND LI 2001)*

***Razón de convergencia**: Bajo ciertas condiciones (entre ellas las condiciones clave 1, 2 y 3 de la Sección 3.3) se tiene que*

$$||\widehat{\boldsymbol{\beta}}_{\boldsymbol{n}} - \boldsymbol{\beta}|| = O_p(n^{-\frac{1}{2}})$$

3.6.2. Número de variables $p = p_n$ tendiendo a infinito cuando n tiende a infinito $(p_n < n)$

En esta sección estudiaremos las propiedades asintóticas del estimador SCAD de mínimos cuadrados penalizados cuando el número de covariables crece con el tamaño de muestra, pero manteniéndose por debajo de n.

Propiedades asintóticas del estimador LS-SCAD

A continuación citaremos las propiedades del estimador penalizado de mínimos cuadrados (SCAD), abreviado como LS-SCAD. Mostraremos que para este estimador LS-SCAD se tiene la consistencia del estimador de los parámetros, la normalidad asintótica de los estimadores de los coeficientes no nulos y la propiedad de identificar correctamente los coeficientes nulos.

El siguiente resultado muestra la consistencia del estimador de los parámetros.

Teorema 3.6.5. *(HUANG AND XIE 2007)*

***Consistencia en la estimación de los parámetros** ($\widehat{\boldsymbol{\beta}}_{\boldsymbol{n}}$):*

Bajo ciertas condiciones (entre ellas las condiciones clave 1, 2, 4 y 8 de la Sección 3.3) se tiene que

$$||\widehat{\boldsymbol{\beta}}_{\boldsymbol{n}} - \boldsymbol{\beta}|| \xrightarrow{p} \text{ cuando } n \to \infty,$$

es decir, el estimador $\widehat{\boldsymbol{\beta}}_{\boldsymbol{n}}$ es consistente.

A continuación daremos otras dos propiedades del estimador LS-SCAD.

Teorema 3.6.6. *(HUANG AND XIE 2007)*

***Normalidad asintótica**: Bajo ciertas condiciones (entre ellas las condiciones clave 1, 2, 4 y 8 de la Sección 3.3), se tiene la normalidad asintótica de los estimadores de los coeficientes no nulos, es decir,*

$$\sqrt{n}\boldsymbol{A_n}(\widehat{\boldsymbol{\beta}}_{\mathbf{1}\boldsymbol{n}} - \boldsymbol{\beta_1}) \to N(\mathbf{0_d}, \boldsymbol{I_d})$$

donde $\boldsymbol{A_n}$ es una matriz de dimensión $d \times k_n$.

Teorema 3.6.7. *(HUANG AND XIE 2007)*

Bajo ciertas condiciones (entre ellas las condiciones clave 1, 2, 4 y 8 de la Sección 3.3) se tiene que

$$P(\widehat{\boldsymbol{\beta}}_{\mathbf{2}\boldsymbol{n}} = \mathbf{0}_{\boldsymbol{m_n}}) \to 1,$$

es decir, con probabilidad 1 se identifican correctamente los coeficientes nulos.

De esta manera, si el número de variables p crece con el tamaño de muestra, para el estimador SCAD se tiene la normalidad asintótica de los coeficientes no nulos y la propiedad de identificar correctamente los coeficientes nulos. Por tanto se dice que el estimador SCAD cumple la propiedad del oráculo, es decir, funciona tan bien como si los coeficientes nulos se conocieran de antemano.

Teorema 3.6.8. *(HUANG AND XIE 2007)*

***Razón de convergencia**: Bajo ciertas condiciones, (entre ellas las condiciones clave 1, 2, 4 y 8 de la Sección 3.3) se tiene que*

$$||\widehat{\boldsymbol{\beta}}_{\boldsymbol{n}} - \boldsymbol{\beta}|| = O_p\left(\sqrt{\frac{p_n}{n}}\right)$$

3.6.3. Número de variables $p = p_n$ tendiendo a infinito cuando n tiende a infinito ($p_n > n$)

Cuando el número de variables aumenta al aumentar el tamaño de muestra se tiene la propiedad de seleccionar correctamente los coeficientes nulos.

Teorema 3.6.9. *(LV y FAN 2009))*

Bajo ciertas condiciones (entre ellas las condiciones clave 1, 2, y 9 de la Sección 3.3 y asumiendo errores normales) se tiene que

$$P(\widehat{\boldsymbol{\beta}}_{\boldsymbol{2n}} = \mathbf{0}_{\boldsymbol{m_n}}) \to 1,$$

es decir, con probabilidad 1 se identifican correctamente los coeficientes nulos.

La siguiente sección estará dedicada a las conclusiones más generales de los métodos de selección de variables.

3.7. Conclusiones

En este capítulo hemos estudiado tres métodos diferentes de selección de variables, el método Lasso, el método Bridge y el método SCAD, tanto cuando el número de variables p es fijo como cuando el número de variables crece a medida que aumenta el tamaño de muestra. Las propiedades deseables que todo criterio de selección debiera satisfacer son la consistencia del estimador de los parámetros, la normalidad asintótica de los estimadores de los coeficientes no nulos y la consistencia en la selección del modelo o, al menos, la propiedad de identificar correctamente los coeficientes nulos con probabilidad tendiendo a 1. Nótese que esta última propiedad es más débil que la de la consistencia en la selección del modelo.

Hemos mostrado resultados que garantizan que el método Lasso verifica la consistencia del estimador de los parámetros y la normalidad asintótica de los estimadores de los coeficientes no nulos (ésta última solo si el número de variables p es fijo), pero no siempre cumple la consistencia en la selección del modelo, ni siquiera la propiedad de identificar correctamente los coeficientes nulos con probabilidad tendiendo a 1. Para conseguir tal consistencia se puede o imponer una condición bastante fuerte, llamada condición de Irrepresentabilidad ó bien modificar el estimador, trabajando por tanto con un nuevo método, el método Lasso Adaptado.

Por otra parte, para el estimador Bridge con parámetro $0 < \gamma < 1$ se garantiza la

consistencia del estimador de los parámetros, la normalidad asintótica de los estimadores de los coeficientes no nulos y la propiedad de identificar correctamente los coeficientes nulos con probabilidad tendiendo a 1 tanto cuando el número de variables p es fijo como cuando el número de variables aumenta a medida que aumenta el tamaño de muestra, pero en este último caso sólo si el número de variables es más pequeño que el tamaño muestral. Por tanto, si el parámetro $0 < \gamma < 1$ y $p < n$ dicho método cumple la propiedad del oráculo. En otro caso, es decir cuando el número de variables es mayor que el tamaño de muestra imponiendo una nueva condición, la condición de ortogonalidad parcial se consigue la consistencia en la selección del modelo y en consecuencia, se verifica también la propiedad de identificar correctamente los coeficientes nulos con probabilidad tendiendo a 1.

Por último, para el método SCAD se tiene la consistencia del estimador de los parámetros, la normalidad asintótica de los estimadores de los coeficientes no nulos y la propiedad de identificar correctamente los coeficientes nulos con probabilidad tendiendo a 1, tanto cuando el número de variables p es fijo como cuando el número de variables aumenta con el tamaño de muestra, pero se mantiene por debajo de él. Por todo ello, se dice que el estimador SCAD cumple la propiedad del oráculo, esto es, dicho estimador funciona tan bien como si los coeficientes nulos se conocieran de antemano. También se ha mostrado que el método SCAD identifica correctamente los parámetros nulos con probabilidad tendiendo a 1 incluso cuando la cantidad de covariables supera al tamaño muestral.

De este modo, parece natural que los métodos que mejor funcionarán son el método SCAD y el método Bridge cuando $0 < \gamma < 1$ y $p < n$, ya que como hemos dicho verifican la propiedad del oráculo, es decir, funcionan tan bien como si los coeficientes nulos se conocieran de antemano.

Hay que tener en cuenta también que cuando el número de variables p_n aumenta a medida que aumenta el tamaño de muestra, pero se mantiene por debajo de n, para que se cumplan las propiedades del método Lasso y del método Bridge se asumió la condición $\frac{p_n}{n} \to 0$ mientras que para el método SCAD se asumió la condición $\frac{p_n^2}{n} \to 0$. Así, el método SCAD permite utilizar un número de variables p_n menor que en el caso de los métodos.

En el siguiente capítulo respaldaremos mediante simulación estos resultados.

Capítulo 4

Comparativa de métodos de selección de variables: Simulación

4.1. Estudio de simulación: Diseño

Realizaremos un estudio de simulación para comparar sobre muestras finitas, el comportamiento de algunos métodos de selección de variables estudiados en el capítulo anterior, tales como el método Lasso, el método Bridge con parámetro $\gamma = 0.5$ y $\gamma = 1.5$ y el método SCAD. En este estudio comparativo también se incluirá el estimador clásico de mínimos cuadrados (por tanto sin penalizar), así como el estimador del ORÁCULO (mínimos cuadrados asumiendo conocidos de antemano cuáles son los coeficientes nulos). Todas las simulaciones están implementadas en R.

Simularemos 200 conjuntos de muestras de tamaño n (los mismos para cada método implementado) del modelo

$$Y = X_1\beta_1 + \ldots + X_p\beta_p + \varepsilon_i \quad \text{para cada } i = 1, \ldots, n,$$

donde $(X_1, \ldots, X_p)$ es un vector normal multivariante de media $\mathbf{0}$ y matriz de varianzas y covarianzas

$$\begin{pmatrix} \rho^{|1-1|} & \rho^{|1-2|} & \dots & \rho^{|1-j|} \dots & \rho^{|1-p|} \\ \rho^{|2-1|} & \rho^{|2-2|} & \dots & \rho^{|2-j|} \dots & \rho^{|2-p|} \\ \vdots & \vdots & \ddots & \vdots & \\ \rho^{|i-1|} & \rho^{|i-2|} & \dots & \rho^{|i-j|} \dots & \rho^{|i-p|} \\ \vdots & \vdots & \ddots & \vdots & \\ \rho^{|p-1|} & \rho^{|p-2|} & \dots & \rho^{|p-j|} \dots & \rho^{|p-p|} \end{pmatrix}$$

donde p es el número de variables y $\rho \in (-1, 1)$. A mayor $|\rho|$ mayor correlación entre las variables.

El vector de parámetros es $\boldsymbol{\beta} = (\beta_1, \beta_2, \dots, \beta_p)' = (1, 2, 0, 2, 0, 0, -4, 0, \dots, 0)'$ y $\varepsilon \in N(0, 1)$.

Para cada $(n, p, \rho) \in \{50, 100, 200\} \times \{10, 75, 150, 300, 1000\} \times \{0, 0.5, 0.9\}$ resumiremos en distintas tablas la siguiente información: (veáse la Sección 4.3)

- Promedio sobre las 200 réplicas de las cantidades de coeficientes no nulos identificados correctamente, seleccionando el parámetro λ por dos criterios, el criterio de validación cruzada generalizada y el criterio BIC.

- Promedio sobre las 200 réplicas de las cantidades de coeficientes nulos identificados correctamente, seleccionando el parámetro λ por dos criterios, el criterio de validación cruzada generalizada y el criterio BIC.

- Promedio sobre las 200 réplicas de la medida de error PMSE seleccionando el parámetro λ por el criterio de validación cruzada generalizada y por el criterio BIC.

  El PMSE se construye de la siguiente manera: Partiendo de una muestra de entrenamiento $\{y_i^e, \boldsymbol{x_i^e}\}_{i=1}^n$ obtenemos el vector de coeficientes estimado $\widehat{\boldsymbol{\beta}}_{\boldsymbol{n}}$ y utilizando una muestra de prueba $\{y_i^p, \boldsymbol{x_i^p}\}_{i=1}^n$, se tiene que

$$PMSE = \frac{\sum_{i=1}^n (y_i^p - \boldsymbol{x_i^{p'}} \widehat{\boldsymbol{\beta}}_{\boldsymbol{n}})^2}{n}.$$

Además mostraremos gráficos relativos a las densidades de los estimadores de los coeficientes

no nulos.

4.2. Código en R

Crearemos una función para calcular tanto el número de coeficientes nulos y no nulos identificados correctamente por el método Lasso, Bridge y SCAD, como la media de la medida de error PMSE de dichos métodos para diferentes tamaños muestrales, diferente número de variables y de correlación entre ellas. Además, dicha función proporcionará los valores estimados del vector de parámetros $\boldsymbol{\beta}$ y el sesgo medio y desviación típica media de los estimadores de $\boldsymbol{\beta}$.

```
general<-function(n,p,M,rho,c.beta,metodo,penalty,crit,gamma){

m1<-matrix(rep(rho,p*p),p,p);
m2<-abs(outer(1:p,1:p,"-"));

vmat<-m1^m2;

b<-matrix(rep(0,(p+1)*M),M,(p+1));

error=rep(0,M);

set.seed(1234)

for (i in 1:M){

X<-rmultnorm(n, matrix(rep(0,p),p,1), vmat);X #muestra de entrenamiento
XP<-rmultnorm(n, matrix(rep(0,p),p,1), vmat);XP #muestra de prueba
XXP<-cbind(rep(1,n),XP);XXP
e<-rnorm(n,0,1)

y<-X%*%c.beta+e;
yp<-XP%*%c.beta+e;

group=1:ncol(X)

metod<-get(metodo)
if(metodo=="grpreg")

{fit <- metod(X, y, group,penalty=penalty);}
```

```
else
{fit<-metod(X,y,group,gamma=gamma)}


#seleccionando lambda por GCV

s<-select(fit,crit=crit);s

b[i,]<-as.vector(s$beta)

error[i]=sum((yp-XXP%*%b[i,])^2)/n
}


#---------------------------
#1. Media del PMSE
#---------------------------


PMSE1=mean(error)


#-------------------------------------------------------------
#2. Sesgo y desviación típica de los betas estimados
#-------------------------------------------------------------


b=b[,-1];b
bn0=which(c.beta!=0);bn0

sesgo=matrix(rep(c.beta[bn0],M),M,length(bn0),byrow=T)-b[,bn0];sesgo
sesgomedio=colMeans(sesgo);sesgomedio

dtipica<-apply(b[,bn0],2,sd)


#-------------------------------------------------------------------------------------
#3. De entre los beta distintos de 0 ¿cuántos se estiman distintos de 0?
#-------------------------------------------------------------------------------------
```

```
bn0=which(c.beta!=0);
length(bn0)

s=rep(0,M);
for(i in 1:M){

l=length(which(b[i,bn0]!=0))

s[i]=l
}
s


K1media=mean(s);K1media


#-----------------------------------------------------------------
#4. De entre los beta que son 0 ¿cuántos se estiman 0?
#-----------------------------------------------------------------


bn0=which(c.beta==0);
length(bn0)


s=rep(0,M);
for(i in 1:M){

l=length(which(b[i,bn0]==0))

s[i]=l
}
s

K2media=mean(s);K2media

list(b,PMSE1,error,sesgomedio,dtipica,K1media,K2media)
}


c.beta<-c(1,2,0,2,0,0,-4,rep(0,(length(p)-7))
M=200
```

```
salidas<-list()
k=0
for (n in c(50,100,200)){
for (p in c(10,75,150,300,1000)){

for (rho in c(0,0.5,0.9)){
for(metodo in c("grpreg","gBridge")){

for (gamma in c(0.5,1.5)){

for (penalty in c("grSCAD","grLasso")){

for (crit in c("GCV","BIC")){
k=k+1

salidas[[k]]<-general(n,p,M,rho,c.beta,metodo,penalty,crit,gamma)
}
}
}
}
}
}
}
```

A continuación programaremos otra función para implementar el método de mínimos cuadrados ordinarios para diferentes tamaños muestrales, diferente número de variables y diferente correlación entre ellas.

```
generallm<-function(n,p,M,rho,c.beta){

m1<-matrix(rep(rho,p*p),p,p);
m2<-abs(outer(1:p,1:p,"-"));

vmat<-m1^m2;vmat #rho^(|i-j|)

b<-matrix(rep(0,(p+1)*M),M,(p+1));

error=rep(0,M)

set.seed(1234)
```

```
for (i in 1:M){

X<-rmultnorm(n, matrix(rep(0,p),p,1), vmat);X #muestra de entrenamiento
XP<-rmultnorm(n, matrix(rep(0,p),p,1), vmat);XP #muestra de prueba
XXP<-cbind(rep(1,n),XP);XXP
e<-rnorm(n,0,1)

y<-X%*%c.beta+e;y
yp<-XP%*%c.beta+e;yp

group=1:ncol(X)

fit<-lm(y~X)
b[i,]<-as.vector(fit$coefficients)

error[i]=sum((yp-XXP%*%b[i,])^2)/n
}


#---------------------------
#1. Media del PMSE
#---------------------------


PMSE1=mean(error)


#-------------------------------------------------------------
#2. Sesgo y desviación típica de los betas estimados
#-------------------------------------------------------------


b=b[,-1];b
bn0=which(c.beta!=0);bn0

sesgo=matrix(rep(c.beta[bn0],M),M,length(bn0),byrow=T)-b[,bn0];sesgo
sesgomedio=colMeans(sesgo);sesgomedio

dtipica<-apply(b[,bn0],2,sd)


#-------------------------------------------------------------------------------
#3. De entre los beta distintos de 0 ¿cuántos se estiman distintos de 0?
#-------------------------------------------------------------------------------
```

```
bn0=which(c.beta!=0);bn0
length(bn0)

s=rep(0,M);s
for(i in 1:M){

l=length(which(b[i,bn0]!=0))
s[i]=l
}
s

K1media=mean(s);K1media


#-----------------------------------------------------------------
#4. De entre los beta que son 0 ¿cuántos se estiman 0?
#-----------------------------------------------------------------


bn0=which(c.beta==0);bn0

s=rep(0,M);s
for(i in 1:M){

l=length(which(b[i,bn0]==0))

s[i]=l
}
s

K2media=mean(s);K2media

list(b,PMSE1,error,sesgomedio,dtipica,K1media,K2media)
}


c.beta<-c(1,2,0,2,0,0,-4,rep(0,(length(p)-7))
M=200
salidaslm<-list()
k=0
for (n in c(50,100,200)){
for (p in c(10,75,150,300,1000)){
```

```
for (rho in c(0,0.5,0.9)){
k=k+1

salidaslm[[k]]<-generallm(n,p,M,rho,c.beta)
}
}
}
```

4.3. Resultados numéricos de la simulación

En las Tablas 4.1 y 4.3 se representa el promedio en las 200 réplicas de las cantidades de los coeficientes no nulos identificados correctamente, seleccionando el parámetro λ por el criterio de validación cruzada generalizada y por el criterio BIC respectivamente.

A continuación en las Tablas 4.2 y 4.4 se representa el promedio en las 200 réplicas de las cantidades de los coeficientes nulos identificados correctamente, seleccionando el parámetro λ por el criterio de validación cruzada generalizada y por el criterio BIC respectivamente.

Por último en las Tablas 4.5 y 4.6 representamos el promedio en las 200 réplicas de la medida de error PMSE, seleccionando de nuevo el parámetro λ por el criterio de validación cruzada generalizada y por el criterio BIC respectivamente.

4.3.1. Tablas

n	p	ρ	SCAD	BR($\gamma = 1.5$)	LASSO	BR($\gamma = 0.5$)	LS	OR
50	10	0	4	4	4	4	4	4
		0.5	3.99	3.99	4	3.99	4	4
		0.9	3.69	3.98	4	3.70	4	4
	75	0	4	3.64	4	3.85	-	4
		0.5	3.99	3.94	4	3.95	-	4
		0.9	3.44	3.40	3.97	2.73	-	4
	150	0	4	3.56	3.99	3.86	-	4
		0.5	3.99	3.89	4	3.94	-	4
		0.9	3.21	3.35	3.96	2.47	-	4
	300	0	4	3.34	4	3.80	-	4
		0.5	4	3.81	4	3.93	-	4
		0.9	3.04	3.40	3.92	2.34	-	4
	1000	0	3.99	3.20	3.97	3.82	-	4
		0.5	3.94	3.83	4	3.92	-	4
		0.9	2.57	3.24	3.52	2.17	-	4
100	10	0	4	4	4	4	4	4
		0.5	4	4	4	4	4	4
		0.9	3.96	4	4	3.97	4	4
	75	0	4	4	4	4	4	4
		0.5	4	4	4	4	4	4
		0.9	3.86	4	4	3.94	4	4
	150	0	4	3.96	4	3.95	-	4
		0.5	4	4	4	4	-	4
		0.9	3.77	3.68	3.99	2.87	-	4
	300	0	4	3.94	4	3.93	-	4
		0.5	4	4	4	3.99	-	4
		0.9	3.70	3.58	4	2.88	-	4
	1000	0	4	3.82	4	3.92	-	4
		0.5	4	4	4	3.98	-	4
		0.9	3.36	3.36	4	2.54	-	4
200	10	0	4	4	4	4	4	4
		0.5	4	4	4	4	4	4
		0.9	4	4	4	3.99	4	4
	75	0	4	4	4	4	4	4
		0.5	4	4	4	4	4	4
		0.9	3.99	3.98	4	3.96	4	4
	150	0	4	4	4	4	4	4
		0.5	4	4	4	4	4	4
		0.9	4	3.98	4	3.97	4	4
	300	0	4	4	4	3.99	-	4
		0.5	4	4	4	4	-	4
		0.9	3.96	3.57	4	3.16	-	4
	1000	0	4	3.99	4	3.97	-	4
		0.5	4	4	4	4	-	4
		0.9	3.90	3.42	4	2.82	-	4

Tabla 4.1: Cantidad media de coeficientes no nulos identificados correctamente seleccionando el parámetro λ por el criterio GCV.

n	p	ρ	SCAD	BR($\gamma = 1.5$)	LASSO	BR($\gamma = 0.5$)	LS	OR
50	10	0	5.46	1.33	3.86	5.62	0	6
		0.5	5.65	1.82	4.10	5.75	0	6
		0.9	5.77	1.84	4.29	5.81	0	6
	75	0	64.73	68.22	62.16	70.76	-	71
		0.5	64.87	67.33	62.80	70.86	-	71
		0.9	67.41	68.00	64.14	70.46	-	71
	150	0	136.17	143.59	132.41	145.78	-	146
		0.5	135.74	142.65	133.17	145.76	-	146
		0.9	137.95	143.24	135.14	144.82	-	146
	300	0	282.96	293.37	277.78	295.74	-	296
		0.5	282.43	292.80	277.81	295.74	-	296
		0.9	281.65	281.33	278.06	294.36	-	296
	1000	0	978.75	976.59	971.21	995.79	-	996
		0.5	977.95	975.75	971.14	995.78	-	996
		0.9	975.61	828.83	963.44	992.64	-	996
100	10	0	5.46	0.81	3.68	5.70	0	6
		0.5	5.69	1.61	3.99	5.76	0	6
		0.9	5.89	1.57	4.18	5.87	0	6
	75	0	60.51	59.48	57.57	65.95	0	71
		0.5	62.52	58.33	60.11	67.45	0	71
		0.9	67.66	61.58	63.67	69.57	0	71
	150	0	140.98	139.96	138.98	146	-	146
		0.5	139.91	136.04	137.77	146	-	146
		0.9	139.94	137.89	135.74	145.23	-	146
	300	0	287.63	290.53	284.20	295.99	-	296
		0.5	285.40	288.25	282.87	295.98	-	296
		0.9	283.58	289.00	278.93	294.91	-	296
	1000	0	979.65	990.69	973.38	995.98	-	996
		0.5	976.56	988.80	971.73	995.95	-	996
		0.9	966.12	955.49	957.03	993.24	-	996
200	10	0	5.30	0.63	3.72	5.60	0	6
		0.5	5.62	1.42	4.19	5.74	0	6
		0.9	5.88	1.21	4.22	5.88	0	6
	75	0	61.38	35.42	57.30	66.83	0	71
		0.5	63.35	36.02	60.04	67.76	0	71
		0.9	68.13	42.08	63.44	69.64	0	71
	150	0	124.81	124.73	121.55	136.46	0	146
		0.5	128.86	123.69	126.07	139.63	0	146
		0.9	140.03	128.60	134.30	143.66	0	146
	300	0	294.63	282.33	293.51	296	-	296
		0.5	293.49	272.17	292.10	296	-	296
		0.9	287.37	279.80	284.20	295.63	-	296
	1000	0	991.91	983.93	988.68	996	-	996
		0.5	988.94	979.94	986.06	996	-	996
		0.9	969.96	981.93	965.74	995.38	-	996

Tabla 4.2: Cantidad media de coeficientes nulos identificados correctamente seleccionando el parámetro λ por el criterio GCV.

n	p	ρ	SCAD	BR($\gamma = 1.5$)	LASSO	BR($\gamma = 0.5$)	LS	OR
50	10	0	4	4	4	4	4	4
		0.5	3.99	3.99	4	3.99	4	4
		0.9	3.82	3.98	4	3.82	4	4
	75	0	4	3.91	4	3.86	-	4
		0.5	3.99	3.98	4	3.96	-	4
		0.9	3.52	3.65	3.98	2.91	-	4
	150	0	4	3.80	3.99	3.86	-	4
		0.5	3.99	3.98	4	3.94	-	4
		0.9	3.25	3.56	3.97	2.69	-	4
	300	0	4	3.65	4	3.80	-	4
		0.5	4	3.97	4	3.93	-	4
		0.9	3.05	3.55	3.99	2.58	-	4
	1000	0	3.99	3.43	3.97	3.82	-	4
		0.5	3.94	3.93	4	3.92	-	4
		0.9	2.57	3.02	3.52	2.31	-	4
100	10	0	4	4	4	4	4	4
		0.5	4	4	4	4	4	4
		0.9	3.99	4	4	3.98	4	4
	75	0	4	4	4	4	4	4
		0.5	4	4	4	4	4	4
		0.9	3.93	4	4	3.97	4	4
	150	0	4	3.97	4	3.95	-	4
		0.5	4	4	4	4	-	4
		0.9	3.80	3.68	3.99	2.92	-	4
	300	0	4	3.96	4	3.93	-	4
		0.5	4	4	4	3.99	-	4
		0.9	3.73	3.58	4	2.93	-	4
	1000	0	4	3.87	4	3.92	-	4
		0.5	4	4	4	3.98	-	4
		0.9	3.36	3.36	4	2.60	-	4
200	10	0	4	4	4	4	4	4
		0.5	4	4	4	4	4	4
		0.9	4	4	4	3.99	4	4
	75	0	4	4	4	4	4	4
		0.5	4	4	4	4	4	4
		0.9	4	3.98	4	3.96	4	4
	150	0	4	4	4	4	4	4
		0.5	4	4	4	4	4	4
		0.9	3.99	3.98	4	3.97	4	4
	300	0	4	4	4	3.99	-	4
		0.5	4	4	4	4	-	4
		0.9	3.97	3.57	4	3.16	-	4
	1000	0	4	3.99	4	3.97	-	4
		0.5	4	4	4	4	-	4
		0.9	3.90	3.42	4	2.82	-	4

Tabla 4.3: Cantidad media de coeficientes no nulos identificados correctamente seleccionando el parámetro λ por el criterio BIC.

n	p	ρ	SCAD	BR($\gamma = 1.5$)	LASSO	BR($\gamma = 0.5$)	LS	OR
50	10	0	4.78	0.82	3.21	5.19	0	6
		0.5	5.19	1.40	3.58	5.36	0	6
		0.9	5.47	1.42	3.98	5.55	0	6
	75	0	63.87	44.81	61.48	70.70	-	71
		0.5	63.79	45.76	61.70	70.84	-	71
		0.9	65.46	54.22	62.52	69.85	-	71
	150	0	136.08	126.51	132.16	145.74	-	146
		0.5	135.52	123.06	132.66	145.71	-	146
		0.9	135.57	134.44	131.83	143.54	-	146
	300	0	282.96	276.96	277.72	295.70	-	296
		0.5	282.43	277.50	277.75	295.58	-	296
		0.9	280.54	283.42	273.81	291.90	-	296
	1000	0	978.75	972.45	971.21	995.79	-	996
		0.5	977.95	974.23	971.14	995.67	-	996
		0.9	975.62	973.19	963.32	989.90	-	996
100	10	0	5.10	0.64	3.36	5.5	0	6
		0.5	5.42	1.40	3.76	5.64	0	6
		0.9	5.79	1.36	4	5.78	0	6
	75	0	56.97	29.31	54.05	63.84	0	71
		0.5	59.67	33.08	56.93	65.50	0	71
		0.9	65.87	40.86	62.43	68.25	0	71
	150	0	140.98	110.65	138.98	146	-	146
		0.5	139.87	106.43	137.77	146	-	146
		0.9	138.61	119.53	134.91	145.10	-	146
	300	0	287.63	273.42	284.20	295.99	-	296
		0.5	285.40	267.71	282.86	295.98	-	296
		0.9	281.68	279.43	277.14	294.71	-	296
	1000	0	979.65	968.15	973.38	995.98	-	996
		0.5	976.56	969.07	971.73	995.98	-	996
		0.9	966.07	972.42	956.76	992.48	-	996
200	10	0	5.23	0.58	3.61	5.53	0	6
		0.5	5.53	1.34	4.05	5.69	0	6
		0.9	5.86	1.15	4.18	5.86	0	6
	75	0	60.35	24.02	56.25	66.31	0	71
		0.5	62.29	26.60	59.09	67.29	0	71
		0.9	67.69	30.16	63.02	69.46	0	71
	150	0	123.34	92.62	119.96	135.87	0	146
		0.5	127.53	100.46	124.57	139.13	0	146
		0.9	139.30	113.14	133.50	143.02	0	146
	300	0	294.63	259.06	293.51	296	-	296
		0.5	293.49	237.36	292.10	296	-	296
		0.9	286.95	256.58	284.15	295.63	-	296
	1000	0	991.91	962.80	988.68	996	-	996
		0.5	988.94	959.94	986.06	996	-	996
		0.9	969.75	968.28	965.62	995.38	-	996

Tabla 4.4: Cantidad media de coeficientes nulos identificados correctamente seleccionando el parámetro λ por el criterio BIC.

n	p	ρ	SCAD	BR($\gamma = 1.5$)	LASSO	BR($\gamma = 0.5$)	LS	OR
50	10	0	1.099	1.387	1.204	1.106	1.264	1.074
		0.5	1.105	1.353	1.192	1.110	1.264	1.078
		0.9	1.176	1.298	1.191	1.174	1.264	1.083
	75	0	1.237	7.547	1.549	1.370	-	1.099
		0.5	1.253	6.863	1.435	1.264	-	1.099
		0.9	1.479	6.256	1.499	1.949	-	1.094
	150	0	1.214	8.353	1.662	1.357	-	1.070
		0.5	1.255	7.658	1.499	1.288	-	1.068
		0.9	1.968	6.456	1.697	2.438	-	1.063
	300	0	1.270	9.425	1.929	1.424	-	1.073
		0.5	1.361	8.727	1.719	1.369	-	1.075
		0.9	2.942	7.009	2.084	3.013	-	1.078
	1000	0	1.306	10.763	2.535	1.419	-	1.074
		0.5	1.489	9.603	2.076	1.415	-	1.084
		0.9	3.763	11.690	3.783	3.821	-	1.099
100	10	0	1.041	1.131	1.081	1.043	1.100	1.035
		0.5	1.038	1.130	1.073	1.042	1.100	1.034
		0.9	1.044	1.112	1.070	1.046	1.100	1.031
	75	0	1.149	3.791	1.193	1.174	4.447	1.035
		0.5	1.138	3.480	1.170	1.133	4.447	1.037
		0.9	1.136	3.744	1.119	1.103	4.447	1.037
	150	0	1.035	5.248	1.246	1.116	-	1.019
		0.5	1.041	4.603	1.189	1.051	-	1.020
		0.9	1.136	4.446	1.230	1.811	-	1.022
	300	0	1.052	6.041	1.298	1.136	-	1.024
		0.5	1.063	5.401	1.244	1.068	-	1.022
		0.9	1.285	5.026	1.328	1.966	-	1.024
	1000	0	1.061	7.002	1.413	1.163	-	1.021
		0.5	1.084	6.181	1.345	1.249	-	1.023
		0.9	1.789	6.165	1.597	2.629	-	1.024
200	10	0	1.025	1.071	1.046	1.026	1.058	1.022
		0.5	1.025	1.070	1.043	1.027	1.058	1.023
		0.9	1.027	1.063	1.042	1.031	1.058	1.023
	75	0	1.032	1.875	1.076	1.525	1.606	1.005
		0.5	1.030	1.794	1.064	1.044	1.606	1.006
		0.9	1.030	2.007	1.073	1.047	1.606	1.007
	150	0	1.078	2.956	1.102	1.126	4.145	1.007
		0.5	1.067	2.757	1.086	1.087	4.145	1.007
		0.9	1.041	3.037	1.093	1.052	4.145	1.007
	300	0	1.012	3.915	1.191	1.058	-	1.018
		0.5	1.013	3.310	1.138	1.038	-	1.010
		0.9	1.054	3.639	1.155	1.640	-	1.011
	1000	0	1.017	4.704	1.193	1.078	-	1.012
		0.5	1.020	4.218	1.144	1.036	-	1.013
		0.9	1.106	4.568	1.215	2.039	-	1.014

Tabla 4.5: Media de la medida de error PMSE seleccionando el parámetro λ por el criterio GCV.

n	p	ρ	SCAD	BR($\gamma = 1.5$)	LASSO	BR($\gamma = 0.5$)	LS	OR
50	10	0	1.108	1.276	1.186	1.120	1.264	1.074
		0.5	1.123	1.288	1.176	1.126	1.264	1.078
		0.9	1.176	1.259	1.178	1.161	1.264	1.083
	75	0	1.244	5.512	1.544	1.392	-	1.099
		0.5	1.255	4.099	1.430	1.267	-	1.099
		0.9	1.700	3.885	1.457	1.982	-	1.094
	150	0	1.216	8.684	1.662	1.363	-	1.070
		0.5	1.255	7.035	1.497	1.305	-	1.068
		0.9	2.167	5.816	1.653	2.589	-	1.063
	300	0	1.270	11.075	1.929	1.436	-	1.073
		0.5	1.361	9.547	1.718	1.401	-	1.075
		0.9	2.988	6.990	2.054	3.367	-	1.078
	1000	0	1.306	13.172	2.535	1.423	-	1.074
		0.5	1.489	11.715	2.076	1.445	-	1.084
		0.9	3.762	9.267	3.873	4.388	-	1.099
100	10	0	1.043	1.113	1.076	1.045	1.100	1.035
		0.5	1.041	1.116	1.069	1.044	1.100	1.034
		0.9	1.046	1.103	1.067	1.048	1.100	1.031
	75	0	1.231	2.210	1.204	1.230	4.447	1.035
		0.5	1.215	2.143	1.179	1.186	4.447	1.037
		0.9	1.197	2.364	1.180	1.137	4.447	1.037
	150	0	1.035	4.565	1.246	1.116	-	1.019
		0.5	1.042	3.189	1.189	1.051	-	1.020
		0.9	1.160	3.308	1.225	1.814	-	1.022
	300	0	1.052	6.466	1.298	1.136	-	1.024
		0.5	1.063	5.321	1.244	1.068	-	1.022
		0.9	1.331	4.860	1.326	1.983	-	1.024
	1000	0	1.061	8.737	1.413	1.163	-	1.021
		0.5	1.084	7.354	1.345	1.125	-	1.023
		0.9	1.791	6.477	1.598	2.707	-	1.024
200	10	0	1.026	1.067	1.046	1.027	1.058	1.022
		0.5	1.026	1.067	1.043	1.027	1.058	1.023
		0.9	1.028	1.060	1.041	1.031	1.058	1.023
	75	0	1.036	1.580	1.076	1.057	1.606	1.005
		0.5	1.036	1.543	1.065	1.049	1.606	1.006
		0.9	1.036	1.586	1.071	1.049	1.606	1.007
	150	0	1.079	2.381	1.104	1.133	4.145	1.007
		0.5	1.073	2.342	1.088	1.094	4.145	1.007
		0.9	1.052	2.622	1.090	1.060	4.145	1.007
	300	0	1.012	3.817	1.191	1.058	-	1.018
		0.5	1.013	2.684	1.138	1.038	-	1.010
		0.9	1.056	2.986	1.155	1.640	-	1.011
	1000	0	1.017	5.202	1.193	1.078	-	1.012
		0.5	1.020	4.464	1.144	1.036	-	1.013
		0.9	1.107	4.609	1.215	2.041	-	1.014

Tabla 4.6: Media de la medida de error PMSE seleccionando el parámetro λ por el criterio BIC

4.3.2. Conclusiones

Cantidad de coeficientes no nulos identificados correctamente

En la Tabla 4.1 se representa la cantidad de coeficientes no nulos identificados correctamente seleccionando el parámetro λ por el criterio de validación cruzada generalizada. Se puede observar que para un tamaño de muestra n fijo, a medida que aumenta el número de variables p se identifican correctamente menos coeficientes no nulos, lo que es natural por el hecho de que al haber más variables es menos probable seleccionar bien las que son no nulas. Por otra parte, si tenemos un tamaño de muestra n fijo y un número de variables p fijo, a medida que aumenta la correlación entre las variables generalmente también se identifican correctamente menos coeficientes no nulos, lo cual es razonable porque a mayor correlación entre las variables menor información se tiene. Además a medida que aumenta el tamaño de muestra n se identifican correctamente más coeficientes no nulos. Esto también es razonable porque a mayor tamaño de muestra se dispone de mayor información. Cabe destacar también que estas conclusiones se tienen sea cual sea el método de selección de variables que utilicemos.

Si comparamos ahora los distintos métodos de selección de variables, se observa que cuando p y ρ son grandes el método Bridge con $\lambda = 0.5$ muestra el peor comportamiento, especialmente si n es pequeño; en el resto de situaciones, los cuatro métodos penalizados arrojan comportamientos parecidos. Como era de esperar, el método de mínimos cuadrados ordinarios identifica como no nulos a todos los coeficientes no nulos.

Si observamos la Tabla 4.3 en la que se representa la cantidad de coeficientes no nulos identificados correctamente pero seleccionando el parámetro λ por el criterio BIC, las conclusiones que obtendríamos serían las mismas. No obstante hemos de destacar que cuando el tamaño de muestra es pequeño ($n = 50$), si seleccionamos el parámetro λ por el criterio BIC, generalmente se identifican correctamente más coeficientes no nulos que si seleccionamos el parámetro λ por el criterio de validación cruzada generalizada. A medida que aumenta el tamaño de muestra (veáse cuando $n = 200$) apenas existen diferencias. En ese caso ambos métodos seleccionan correctamente el mismo número de coeficientes no nulos.

Cantidad de coeficientes nulos identificados correctamente

El efecto de la dependencia se observa especialmente cuando se pasa de dependencia nula o moderada ($\rho = 0, 0.5$) a dependencia alta ($\rho = 0.9$), siendo además el número de variables p muy grande. En estas situaciones, los métodos de selección de variables funcionan claramente mejor cuando la dependencia es nula o moderada. El método menos afectado por la dependencia parece ser el método Bridge con parámetro $\lambda = 0.5$.

Si comparamos los distintos métodos penalizados, se puede observar también que el estimador que mejor identifica correctamente los coeficientes nulos es el estimador Bridge con $\gamma = 0.5$ seguido del estimador SCAD. Recordemos que en teoría habíamos visto que el estimador SCAD y el estimador Bridge cuando $0 < \gamma < 1$ tenían la propiedad de identificar correctamente los coeficientes nulos con probabilidad tendiendo a 1. El estimador Lasso funciona también bastante bien, y el que peor identifica correctamente los coeficientes nulos es el estimador Bridge con $\gamma = 1.5$, lo cual también es razonable ya que en teoría se había visto que el estimador Bridge funcionaba bien cuando el parámetro γ era menor que 1. El estimador de mínimos cuadrados ordinarios no identifica ningún coeficiente a cero. Esto es esperable pues este estimador como se había visto en teoría puede estimar coeficientes muy próximos a cero pero sin llegar a ser exactamente cero.

Como referencia también aparecen los resultados de aplicar mínimos cuadrados ordinarios asumiendo que conocemos de antemano los coeficientes nulos (ORÁCULO). Por eso, en ese caso, en la columna correspondiente a dicho método se representa exactamente la cantidad de coeficientes nulos para cada caso.

En la Tabla 4.4 se representa la cantidad de coeficientes nulos identificados correctamente, seleccionando el parámetro λ por el criterio BIC. Las conclusiones que obtendríamos serían las mismas que si se selecciona el parámetro λ por el criterio de validación cruzada generalizada. Hemos de destacar que con el criterio de selección BIC por lo general, se identifican correctamente menos coeficientes nulos.

Media del PMSE

En la Tabla 4.5 se representa la media de la medida de error PMSE seleccionando el

parámetro λ por el criterio de validación cruzada generalizada. Nótese que un método que se comporte bien debería llevarnos a un valor del PMSE próximo a 1 pues es la varianza del error ($\varepsilon \sim N(0,1)$).

El efecto de la dependencia se observa especialmente cuando se pasa de dependencia nula o moderada ($\rho = 0, 0.5$) a dependencia alta ($\rho = 0.9$), siendo además el número de variables p muy grande. En estas situaciones, los métodos de selección de variables funcionan claramente mejor cuando la dependencia es nula o moderada. Además para un tamaño de muestra n fijo, a medida que aumentamos el número de variables, los errores también se van haciendo cada vez más grandes. Esto también es lo esperado puesto que los coeficientes estimados también van a ser menos precisos y en consecuencia se tendrá un error mayor en la predicción. Cabe destacar también que según aumenta el tamaño de muestra, los errores van disminuyendo y aproximándose cada vez más a 1. Esto es esperable puesto que a mayor tamaño de muestra mayor precisión a la hora de estimar los coeficientes y por tanto menor error en la predicción.

Por otra parte, si comparamos los distintos métodos, podemos observar que el método Bridge con parámetro $\gamma = 1.5$ proporciona unos errores muy grandes debido a que dicho método con un parámetro mayor que 1 no funciona bien. A continuación le sigue el estimador Lasso. Por último nótese que el estimador de mínimos cuadrados ordinarios proporciona unos errores altos, pues dicho estimador no identifica correctamente los coeficientes nulos (no fija a 0) y por tanto el error en la predicción será mayor. El estimador que menor error proporciona naturalmente es el del oráculo, pues conoce de antemano los coeficientes nulos.

La Tabla 4.6 proporciona la media de la medida de error de PMSE pero seleccionando el parámetro λ por el criterio BIC. Las conclusiones obtenidas son las mismas que si seleccionamos el parámetro λ por el criterio de validación cruzada generalizada.

Generalmente se obtienen unos errores un poco más pequeños seleccionando el parámetro λ por el criterio GCV. Estos errores se van igualando a medida que el tamaño de muestra aumenta. El hecho de que se obtengan unos errores más pequeños con el criterio GCV está relacionado con que con dicho criterio se identifican correctamente mejor los coeficientes nulos y por tanto menor error habrá en la predicción.

4.4. Representaciones gráficas

En esta sección vamos a representar gráficamente las estimaciones de las funciones de densidad de los estimadores de los coeficientes no nulos obtenidos utilizando los dos métodos de selección de variables que parecen dar mejores resultados tanto desde el punto de vista teórico como aplicado (veánse las secciones 3.7 y 4.3.2, respectivamente): el método Bridge con parámetro $\gamma = 0.5$ y el método SCAD. Además a modo de referencia, incluiremos las densidades estimadas correspondientes al estimador del oráculo. Representaremos las estimaciones de las funciones de densidad para distintos tamaños de muestra $n = 50, 100, 200$, distinto nº de variables $p = 10, 150, 1000$ y distinta correlación entre las variables $\rho = 0, 0.9$.

Utilizaremos el estimador tipo núcleo para estimar las densidades que viene dado por

$$\widehat{f}_{n,k} = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right) = \frac{1}{n}\sum_{i=1}^{n} K_h(x - X_i)$$

donde $K_h(u) = \frac{1}{h}K(u/h)$. La función núcleo es, generalmente, una función de densidad simétrica con respecto al origen, como por ejemplo, la densidad de la $N(0, 1)$. La selección del parámetro ventana h es clave: si h toma valores muy pequeños, tenderemos a infrasuavizar, y si h toma valores excesivamente grandes, la estimación estará sobresuavizada.

Para evaluar globalmente el comportamiento de $\widehat{f}_{n,k}$, como estimador de f, resulta conveniente utilizar una medida de error global, como el $MISE$, dada por:

$$MISE(h) = \frac{1}{nh}R(K) + \frac{1}{4}h^4\mu_2(K)^2R(f'') + o((nh)^{-1} + h^4),$$

donde $R(K) = \int K^2(u)du$ y $\mu_2(K) = \int u^2K(u)du$. La aproximación asintótica, el $AMISE$, viene dada por:

$$AMISE(h) = \frac{1}{nh}R(K) + \frac{1}{4}h^4\mu_2(K)^2R(f'').$$

Por tanto, simplemente minimizando en h podemos ver que la ventana óptima resultante es:

$$h_{AMISE} = \left(\frac{R(K)}{\mu_2(K)^2R(f'')n}\right)^{1/5}$$

4.4.1. Densidades estimadas

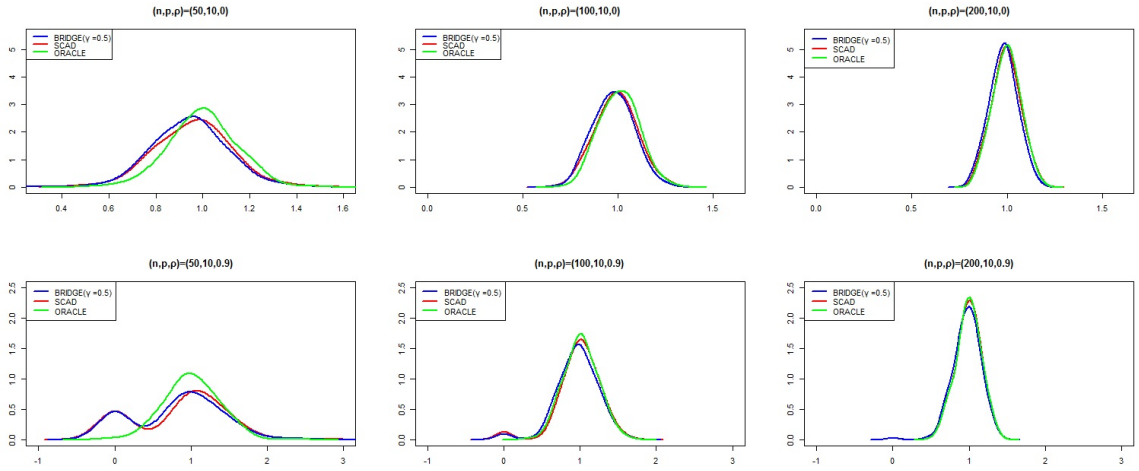

Figura 4.1: Representación gráfica de la función de densidad de $\widehat{\beta}_1$ en modelos con $p = 10$ variables utilizando diferentes tamaños muestrales, diferente correlacción entre las variables, y diferentes métodos de estimación de tal coeficiente.

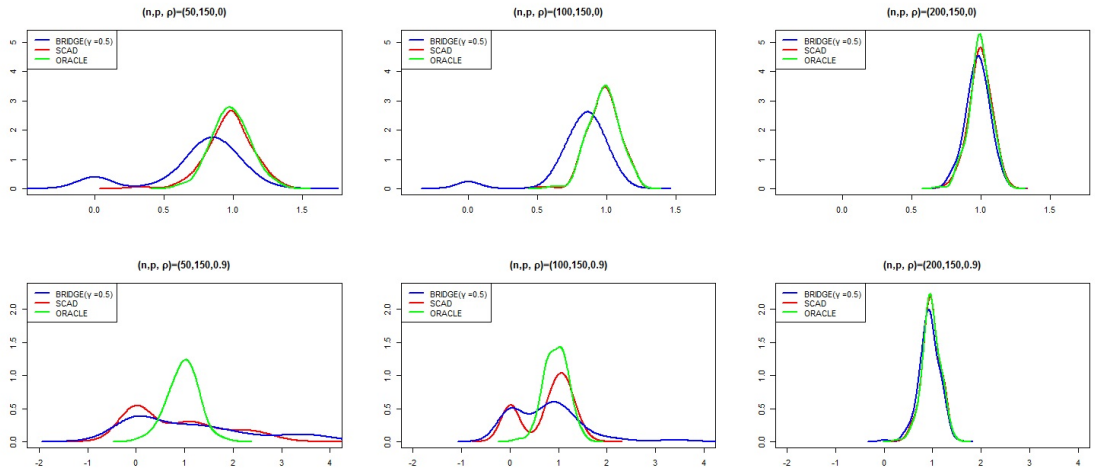

Figura 4.2: Representación gráfica de la función de densidad de $\widehat{\beta}_1$ en modelos con $p = 150$ variables utilizando diferentes tamaños muestrales, diferente correlacción entre las variables, y diferentes métodos de estimación de tal coeficiente.

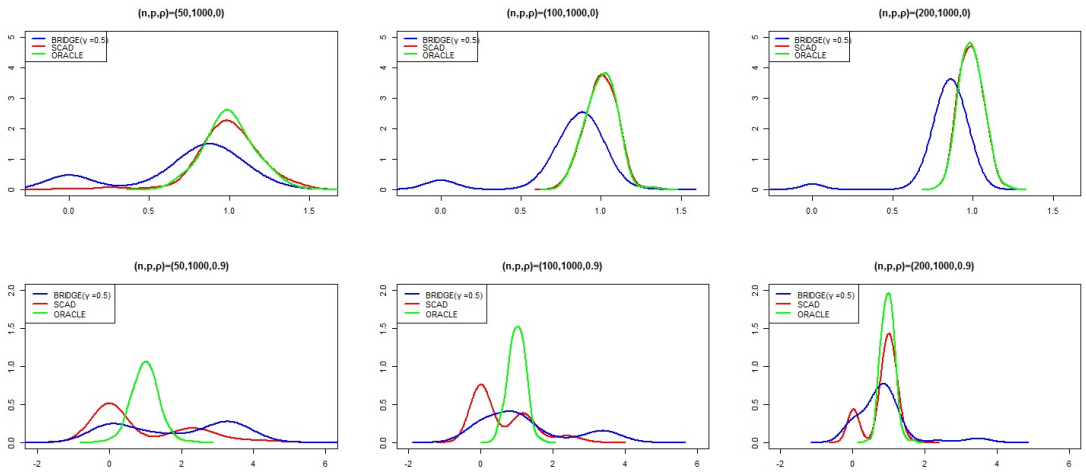

Figura 4.3: Representación gráfica de la función de densidad de $\widehat{\beta_1}$ en modelos con $p = 1000$ variables utilizando diferentes tamaños muestrales, diferente correlacción entre las variables, y diferentes métodos de estimación de tal coeficiente.

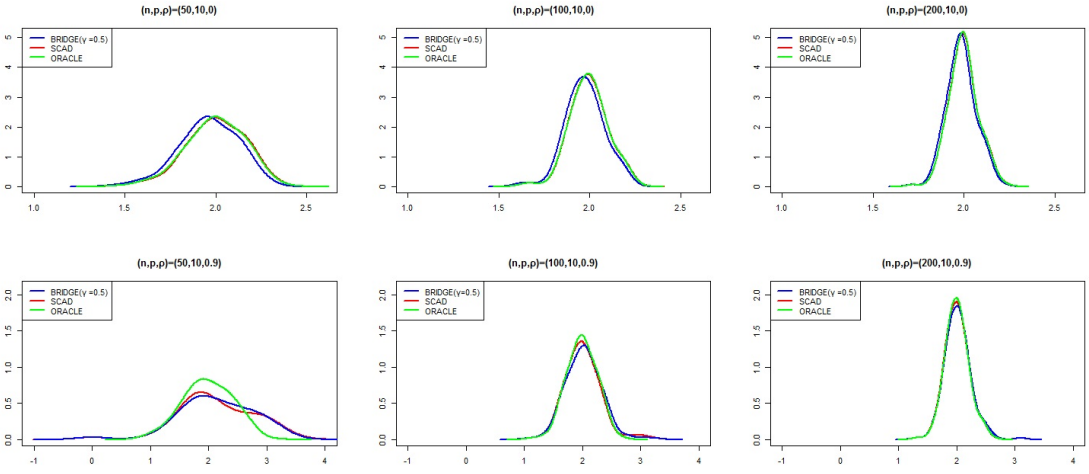

Figura 4.4: Representación gráfica de la función de densidad de $\widehat{\beta_2}$ en modelos con $p = 10$ variables utilizando diferentes tamaños muestrales, diferente correlacción entre las variables, y diferentes métodos de estimación de tal coeficiente.

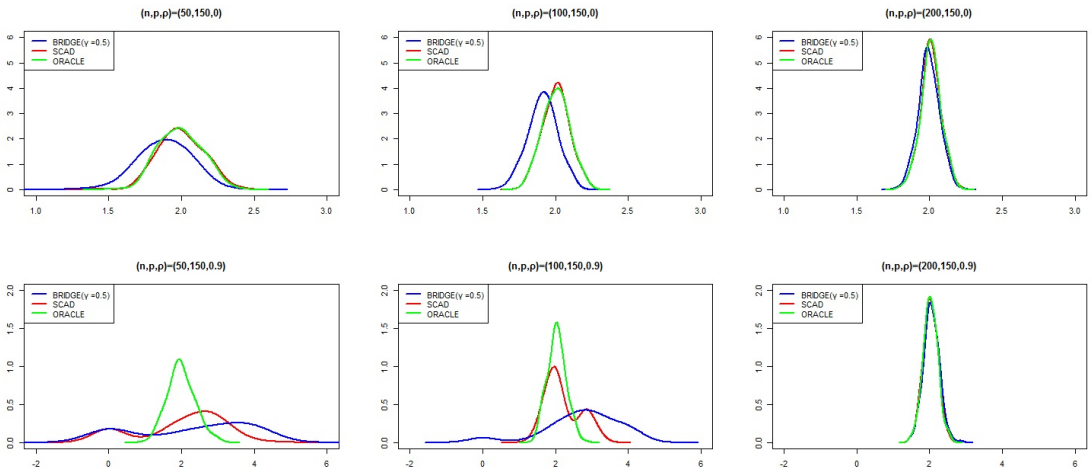

Figura 4.5: Representación gráfica de la función de densidad de $\widehat{\beta}_2$ en modelos con $p = 150$ utilizando diferentes tamaños muestrales, diferente correlacción entre las variables, y diferentes métodos de estimación de tal coeficiente.

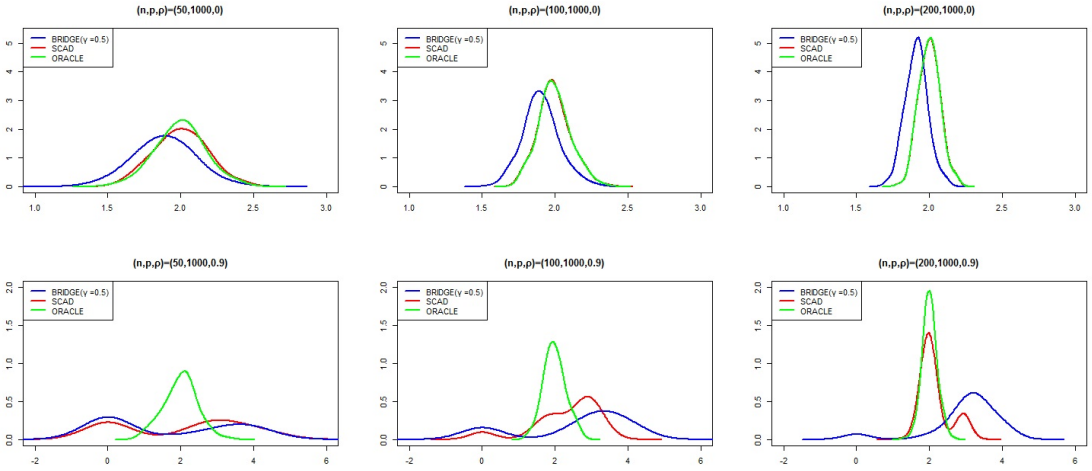

Figura 4.6: Representación gráfica de la función de densidad de $\widehat{\beta}_2$ en modelos con $p = 1000$ variables utilizando diferentes tamaños muestrales, diferente correlacción entre las variables, y diferentes métodos de estimación de tal coeficiente.

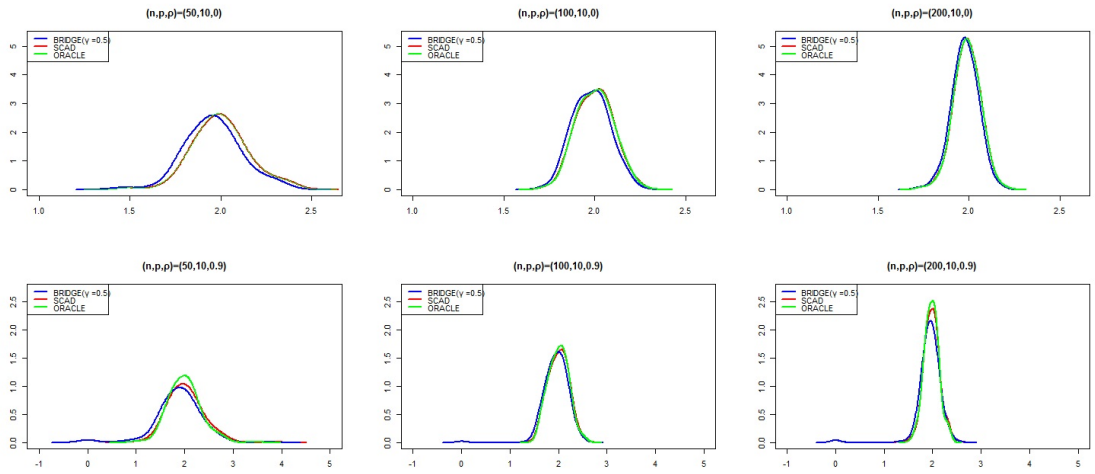

Figura 4.7: Representación gráfica de la función de densidad de $\widehat{\beta}_4$ en modelos con $p = 10$ variables utilizando diferentes tamaños muestrales, diferente correlacción entre las variables, y diferentes métodos de estimación de tal coeficiente.

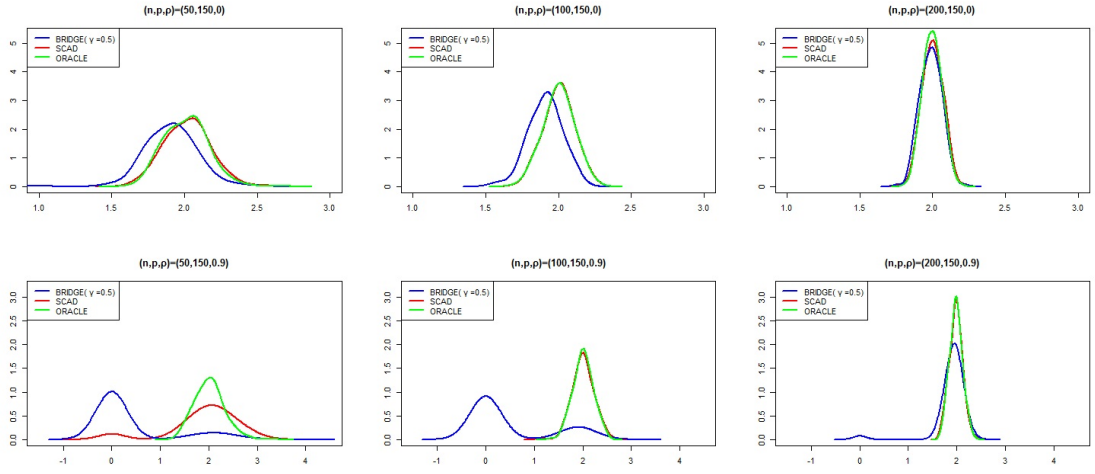

Figura 4.8: Representación gráfica de la función de densidad de $\widehat{\beta}_4$ en modelos con $p = 150$ variables utilizando diferentes tamaños muestrales, diferente correlacción entre las variables, y diferentes métodos de estimación de tal coeficiente.

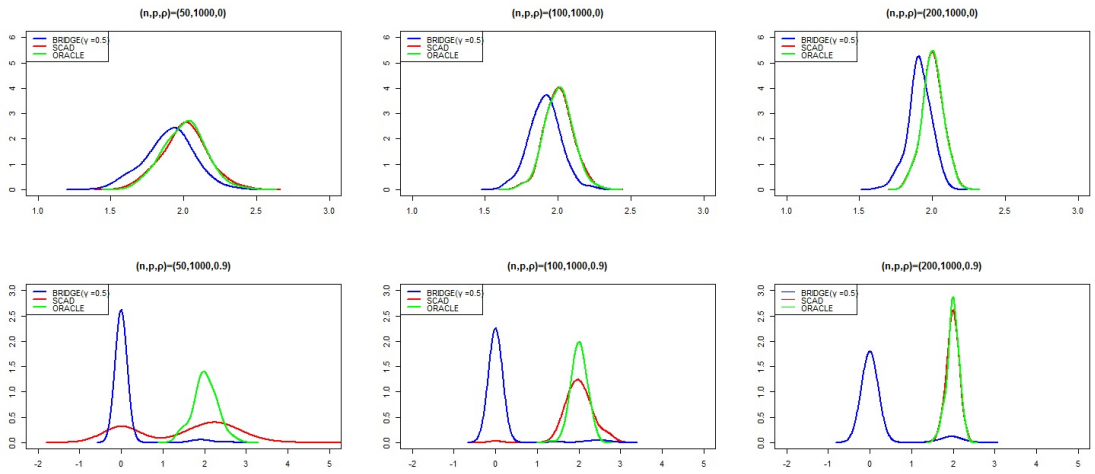

Figura 4.9: Representación gráfica de la función de densidad de $\widehat{\beta_4}$ en modelos con $p = 1000$ variables utilizando diferentes tamaños muestrales, diferente correlacción entre las variables, y diferentes métodos de estimación de tal coeficiente.

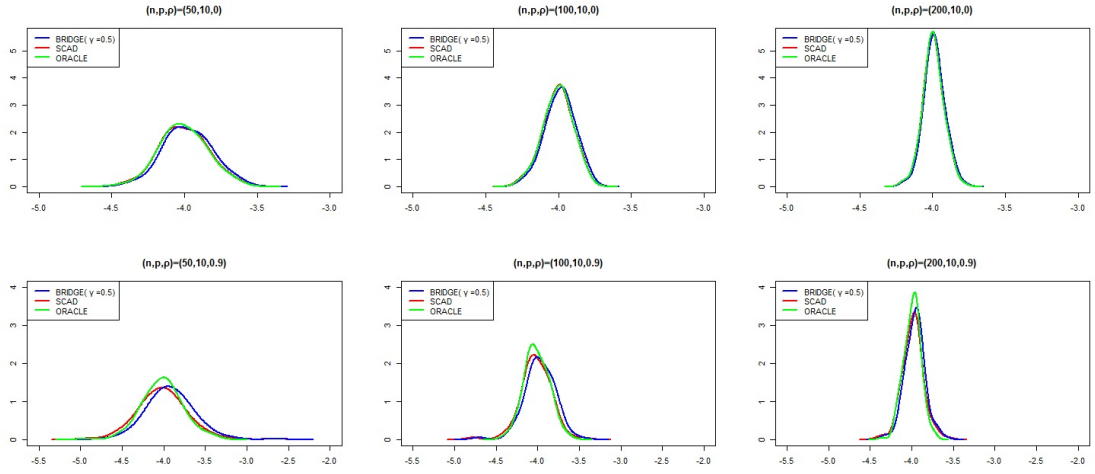

Figura 4.10: Representación gráfica de la función de densidad de $\widehat{\beta_7}$ en modelos con $p = 10$ variables utilizando diferentes tamaños muestrales, diferente correlacción entre las variables, y diferentes métodos de estimación de tal coeficiente.

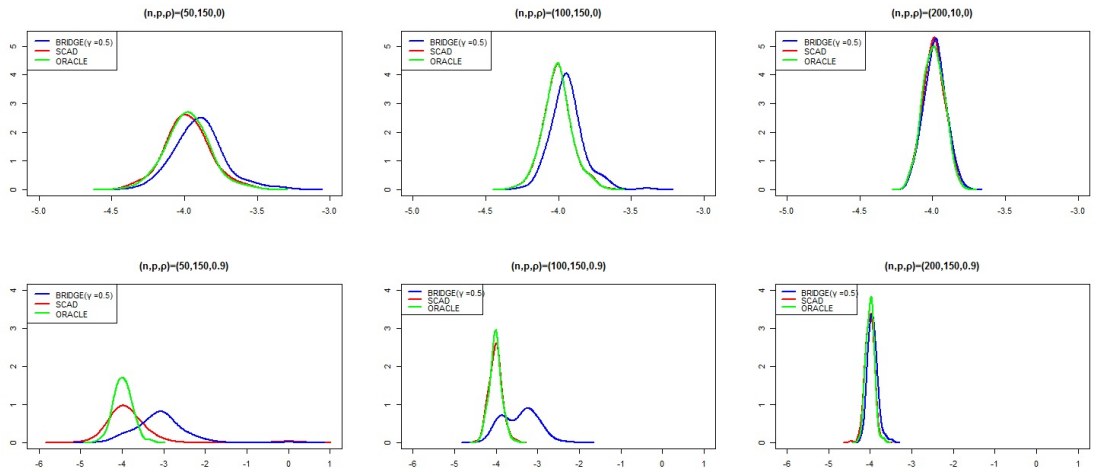

Figura 4.11: Representación gráfica de la función de densidad de $\widehat{\beta_7}$ en modelos con $p = 150$ variables utilizando diferentes tamaños muestrales, diferente correlacción entre las variables, y diferentes métodos de estimación de tal coeficiente.

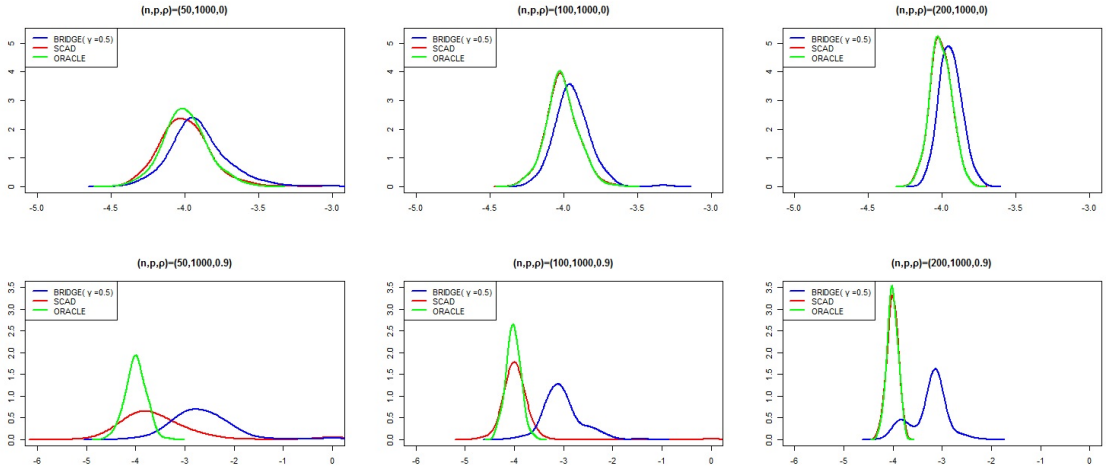

Figura 4.12: Representación gráfica de la función de densidad de $\widehat{\beta_7}$ en modelos con $p = 1000$ variables utilizando diferentes tamaños muestrales, diferente correlacción entre las variables, y diferentes métodos de estimación de tal coeficiente.

4.4.2. Conclusiones

Como hemos visto en teoría los coeficientes estimados se distribuyen asintóticamente según una distribución normal. Por tanto, cabe esperar que las densidades de tales coeficientes se asemejen a dicha distribución. Representamos las densidades estimadas de las estimaciones de los 4 coeficientes no nulos, utilizando el método SCAD y el método Bridge con $\gamma = 0.5$ (recuérdese que ambos métodos son los que presentan mejores resultados). Además, también mostramos las correspondientes al estimador del Oráculo. Por tanto, con este método las densidades de los coeficientes se aproximarán a la de una distribución normal, y nos servirá como referencia para comprobar si los estimadores de los coeficientes no nulos estimados mediante el método SCAD y Bridge, tienen la misma distribución asintótica que tendrían si los coeficientes nulos fueran conocidos de antemano. Cuanto más se aproximen las densidades estimadas de los coeficientes utilizando el método SCAD o Bridge a la densidad estimada utilizando el método del Oráculo mejor comportamiento del método.

Podemos observar en las gráficas que cuando tanto el tamaño de muestra n y el número de variables p es fijo, a medida que aumenta la correlación entre las variables, las densidades estimadas de las estimaciones de los coeficientes parecen diferir más de la de una distribución normal y también las densidades estimadas obtenidas utilizando el estimador SCAD y Bridge se van diferenciando más de la del Oráculo. No obstante a medida que el tamaño de muestra n aumenta estas diferencias se hacen cada vez más pequeñas (comparemos primera y segunda fila de cada gráfica). A modo de ejemplo se puede ver en la Figura 4.2, que cuando $\rho = 0.9$ la densidad estimada del coeficiente estimado $\widehat{\beta}_1$ utilizando el método SCAD o el método Bridge se asemeja menos a una distribución normal que si $\rho = 0$. Se observa también que a medida que n aumenta, por ejemplo cuando $n = 200$ las densidades son muy similares a la densidad de una distribución normal y tanto la densidad estimada obtenida utilizando el estimador SCAD como la obtenida utilizando el estimador Bridge se asemeja a la del Oráculo. Esto ocurre para todos los coeficientes estimados.

Por otra parte, cuando el tamaño de muestra es fijo y aumentamos el número de variables las densidades estimadas de las estimaciones de los coeficientes parecen distinguirse más de la de una distribución normal y también ocurre que las diferencias entre las densidades

estimadas utilizando el estimador SCAD ó Bridge y el del Oráculo son más acusadas. Por ejemplo, para el coeficiente $\widehat{\beta}_1$ se puede observar en las Figuras 4.1 y 4.3 que cuando n es fijo y $p = 10$ las densidades estimadas de tal coeficiente se asemejan más a la de una distribución normal que cuando $p = 1000$ y además las densidades estimadas obtenidas mediante el método SCAD y Bridge son más parecidas a la del Oráculo cuando $p = 10$. Se observa también que el método SCAD parece funcionar mejor que el método Bridge, ya que la densidad estimada de los coeficientes por el método SCAD se aproxima bastante más a la del oráculo que la estimada por el método Bridge. Estas diferencias son más notables cuando el número de variables es grande ($p = 150$ ó $p = 1000$).

Por último cabe destacar también que a medida que el tamaño de muestra aumenta las densidades estimadas de las estimaciones de los coeficientes se van aproximando más a la de una distribución normal y como se observa las densidades estimadas utilizando el estimador SCAD y Bridge son cada vez más parecidas a la densidad utilizando el método del Oráculo. Veáse por ejemplo en la Figura 4.2 que cuando $n = 50$ y $\rho = 0.9$ las densidades estimadas de las estimaciones de los coeficientes obtenidas mediante el método SCAD y Bridge se distinguen bastante de la del Oráculo. Sin embargo a medida que n crece estas diferencias son menos acusadas.

Capítulo 5

Comparativa de métodos de selección de variables: Aplicación a datos reales

5.1. La base de datos y la aplicación

En este ejemplo estudiamos un conjunto de datos que contiene variables que influyen en la diabetes. Dicho conjunto de datos se ha sacado de

`http://www4.stat.ncsu.edu/~boos/var.select/diabetes.html`

y contiene 10 variables estandarizadas: age (edad), sex (sexo), bmi (índice de masa corporal), bp (promedio de la presión arterial), y 6 medidas de suero sanguíneo: tc (hormona triacontanol), ldl (colesterol-LDL), hdl (colesterol-HDL), tch (hormona tiroidea), ltg (molécula lamotrigina) y glu (glucosa) observadas sobre 442 pacientes con diabetes. La variable respuesta será una medida cuantitativa de la progresión de la enfermedad un año después del inicio del estudio. La muestra se dividirá en dos submuestras: una muestra de entrenamiento formada por las observaciones tomadas sobre los primeros 300 individuos, y una muestra de prueba procedente de los restantes 142 individuos. La primera se utilizará para poner en práctica diversos métodos de selección de variables, mientras que la segunda servirá para

cuantificar su potencia predictiva. Implementaremos el método Lasso, el método Bridge con parámetro $\gamma = 0.5$ y el método SCAD. Seleccionaremos el parámetro λ por el criterio de validación cruzada generalizada y por el criterio BIC. En la siguiente sección mostraremos el código R utilizado para la implementación, mientras que en la Sección 5.3 se presentarán los resultados.

5.2. Código en R

```
#------------------------
#Cargamos librerías
#------------------------


library(grpreg)


#------------------------
#Lectura de datos
#------------------------


d<-read.table("datos.txt",header=T)


#muestra de entrenamiento


X<-as.matrix(d[1:300,1:10])
XX<-as.matrix(cbind(rep(1,300),X));XX


#muestra de prueba


XP<-as.matrix(d[301:442,1:10])
XXP<-as.matrix(cbind(rep(1,142),XP));XXP


#respuesta en la muestra de entrenamiento


y<-d[1:300,11]
```

```
#respuesta en la muestra de prueba


yp<-d[301:442,11]


#---------------------
#Método SCAD
#---------------------


group=1:ncol(X)
fit <- grpreg(X, y, group,penalty="grSCAD");fit
summary(fit)


#seleccionando lambda por GCV


s<-select(fit,crit="GCV");s


#-------------------------------------
#1. Coeficientes beta estimados
#-------------------------------------


b<-as.vector(s$beta);b


#-----------------------------------------
#2. Potencia de predicción
#-----------------------------------------


error=sum((yp-XXP%*%b)^2)/length(yp);error


#---------------------------------------
#3.R^2 ajustado
#---------------------------------------
#Suma de cuadrados residual


RSS=sum((y-XX%*%b)^2)/(300-11)
```

```
#Suma de cuadrados total

TSS=sum((y-mean(y))^2)/(300-1)

r2ajustado=1-(RSS/TSS);r2ajustado

#---------------------
#Método Lasso
#---------------------

fit <- grpreg(X, y, group,penalty="grLasso");fit
summary(fit)

#seleccionando lambda por GCV

s<-select(fit,crit="GCV");s

#----------------------------------------
#1. Coeficientes beta estimados
#----------------------------------------

b<-as.vector(s$beta);b

#----------------------------------------
#2. Potencia de predicción
#----------------------------------------

error=sum((yp-XXP%*%b)^2)/length(yp);error

#----------------------------------------
#3.R^2 ajustado
#----------------------------------------
#Suma de cuadrados residual

RSS=sum((y-XX%*%b)^2)/(300-11)
```

```
#Suma de cuadrados total

TSS=sum((y-mean(y))^2)/(300-1)

r2ajustado=1-(RSS/TSS);r2ajustado

#---------------------
#Método Bridge
#---------------------

fit <- gBridge(X, y, group,gamma=0.5);fit

#seleccionando lambda por GCV

s<-select(fit,crit="GCV");s

#---------------------------------------
#1. Coeficientes beta estimados
#---------------------------------------

b<-as.vector(s$beta);b

#---------------------------------------
#2. Potencia de predicción
#---------------------------------------

error=sum((yp-XXP%*%b)^2)/length(yp);error

#---------------------------------------
#3.R^2 ajustado
#---------------------------------------
#Suma de cuadrados residual

RSS=sum((y-XX%*%b)^2)/(300-11)

#Suma de cuadrados total

TSS=sum((y-mean(y))^2)/(300-1)

r2ajustado=1-(RSS/TSS);r2ajustado
```

5.3. Resultados de la aplicación

En esta sección se presentarán los resultados correspondientes a la estimación de los parámetros asociados a las distintas variables, el R^2 ajustado y la media del error cuadrático de predicción para el método Lasso, el método Bridge con parámetro $\gamma = 0.5$ y el método SCAD. La Tabla 5.1 presenta estos resultados cuando se selecciona el parámetro λ por el criterio GCV y la Tabla 5.2 cuando se selecciona el parámetro λ por el criterio BIC.

Método \ Variables	**age**	**sex**	**bmi**	**bp**	**tc**	**ldl**
LASSO	0	0	5.276	0.296	0	0
SCAD	0	0	7.577	0	0	0
BRIDGE	0	0	6.431	0	0	0

	hdl	**tch**	**ltg**	**glu**	$\boldsymbol{R^2}$	**PMSE**
LASSO	-0.104	0	43.652	0	0.4307	3061.207
SCAD	0	0	45.054	0	0.4354	3120.636
BRIDGE	0	0	53.627	0	0.4384	3131.742

Tabla 5.1: Coeficientes estimados, R^2 ajustado y PMSE para los distintos métodos seleccionando el parámetro por GCV.

Método \ Variables	**age**	**sex**	**bmi**	**bp**	**tc**	**ldl**
LASSO	0	-12.917	5.787	0.708	0	-0.080
SCAD	0	-7.696	6.727	0.712	-0.208	0
BRIDGE	0	-15.998	6.354	0.847	-0.328	0

	hdl	**tch**	**ltg**	**glu**	$\boldsymbol{R^2}$	**PMSE**
LASSO	-0.636	0	46.513	0.234	0.4842	2801.456
SCAD	-0.144	0	60.902	0	0.4809	2874.289
BRIDGE	0	5.667	57.202	0	0.4908	2804.402

Tabla 5.2: Coeficientes estimados, R^2 ajustado y PMSE para los distintos métodos seleccionando el parámetro por el criterio BIC

5.4. Conclusiones

Seleccionando el parámetro λ por el criterio GCV.

Con el método SCAD y con el método Bridge seleccionamos las variables bmi y ltg, las demás variables corresponden a coeficientes que se fijan a 0 (se fijan a cero 8 coeficientes). Sin embargo con el método Lasso se seleccionan además de las variables bmi y ltg, las variables bpi y hdl (fija a cero 6 coeficientes). No se aprecian diferencias significativas ni en la bondad del ajuste (R^2) ni en la potencia predictiva (PMSE).

Seleccionando el parámetro λ por el criterio BIC.

Con el método SCAD se seleccionan las variables sex, bmi, bp, tc, hdl y ltg (fija a cero 4 coeficientes). Por otra parte con el método Bridge se seleccionan las variables sex, bmi, bp, tc, tch y ltg (fija a cero 4 coeficientes). Y con el método Lasso se seleccionan las variables sex, bmi, bp, ldl, hdl, ltg y glu (fija a cero 3 coeficientes). No se aprecian diferencias significativas ni en la bondad del ajuste (R^2) ni en la potencia predictiva (PMSE).

Los errores son más pequeños seleccionando el parámetro λ por el criterio BIC y también

proporcionan un R^2 ajustado más alto para todos los métodos que usando el criterio GCV para seleccionar el parámetro λ. Sin embargo, los modelos que utilizan el parámetro λ seleccionando a través del criterio GCV resultan mucho más sencillos puesto que como hemos visto seleccionan bastantes menos variables.

A continuación representaremos la media de los errores de predicción para los distintos métodos estudiados seleccionando el parámetro λ tanto por el criterio GCV como por el criterio BIC.

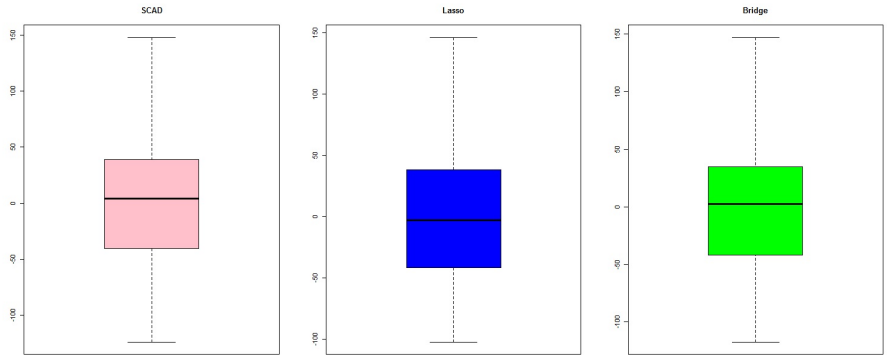

Figura 5.1: Media de los errores de predicción para los métodos SCAD, Lasso y Bridge seleccionando el parámetro λ por el criterio GCV.

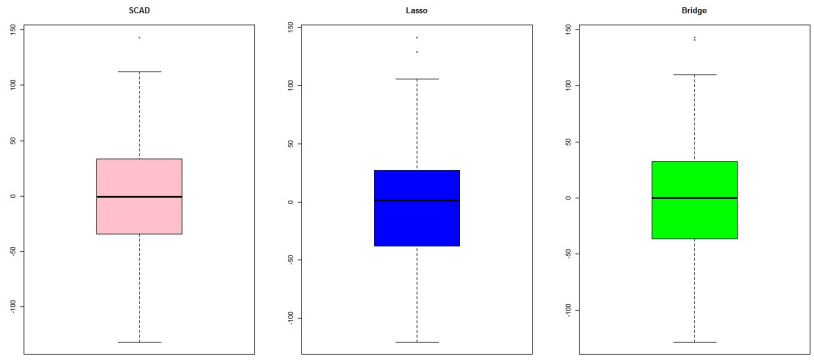

Figura 5.2: Media de los errores de predicción para los métodos SCAD, Lasso y Bridge seleccionando el parámetro λ por el criterio BIC.

Las Figuras 5.1 y 5.2 muestran diagramas de cajas correspondientes a los errores de predicción obtenidos sobre la muestra de entrenamiento cuando se aplican los tres métodos penalizados con cada uno de los selectores de λ. Fijado el selector de λ, no se observan grandes diferencias entre ellos. Comparando ambas figuras, parece que el selector BIC tiende a generar errores más concentrados, lo que redunda en un menor PMSE. También es destacable la presencia de unos pocos atípicos cuando se utiliza tal selector.

REFERENCIAS

[1] Fan, J., and Li, R. (2001). "Variable selection via noncave penalized likelihood and its oracle properties ", *Journal of the American Statistical Association*, 96,1348-1360.

[2] Frank, I., and Friedman, J. (1993). "A statistical view of some chemometries regression tools ", *Technometrics*, 35, 109-148.

[3] Fu, W.J., (1998). "Penalized regressions: The bridge versus the lasso ", *Journal of Computational and Graphical Statistics*, 7, 397-416.

[4] Huang, J., Horowitz, J.L., and Ma, S. (2008). "Asymptotic properties of bridge estimators in sparse high-dimensional regression models ", *The Annals of Statistics*, 36, 587-613.

[5] Huang, J., and Xie, H. (2007). "Asymptotic oracle properties of SCAD-penalized least squares estimators ", *IMS Lecture Notes-Monograph Series Asymptotic: Particles, Processes and Inverse Problems*, 55, 149-166.

[6] Knight, K., and Fu, W. (2000). "Asymptotics for lasso-type estimators ", *The Annals of Statistics*, 28, 1356-1378.

[7] Lv, J., and Fan, Y. (2009). "A unified approach to model selection and sparse recovery using least squares " *The Annals of Statistics*, 37, 3498-3528.

[8] Tibshirani, R. (1996) "Regression shrinkage and selection via the lasso ", *Journal of the Royal Society*, Ser. B, 58, 267-288.

[9] Zhao, P., and Yu, B. (2006). "On model selection consistency of lasso ", *Journal of Machine Learning Research*, 7, 2541-2563.

[10] Zou, H. (2006). "The adaptive lasso and its oracle properties ", *Journal of the*

American Statistical Association, 101, 1418-1429.

Índice de figuras