Facultad de Ciencias Económicas y Estadística

Universidad Nacional de Rosario

Regresión Lineal Múltiple en Grandes Dimensiones

Carrera: Licenciatura en Estadística

Alumno:

Iván Ariel Millanes

Directora:

Dra. Marta Quaglino

Co-Directora:

Lic. María Belén Allasia

2017

Agradecimientos

A mis padres, Silvana y Ariel, por su apoyo incondicional y por brindarme todas las facilidades para transitar esta etapa.

A mi hermano, Ramiro, por bancarme siempre.

A María Rosa, por su cariño incondicional.

A Marta, mi directora, por su paciencia y sabiduría.

A Belén, mi co-directora, por ayudarme a mejorar constantemente y estar siempre dispuesta a darme una mano.

A mis compañeros, en especial a Yami, por transitar este camino juntos.

A mis amigos, Beto, José, Martín y Román, por acompañarme todos estos años.

A los profesores, de quienes siempre aprendí mucho.

Índice general

Capítulo 1

Introducción

El análisis de regresión es una técnica estadística utilizada para investigar y modelar la relación entre variables. Esta técnica estudia la relación entre una variable respuesta o dependiente y una o más variables explicativas o predictores. Se puede usar con un fin descriptivo, es decir, para conocer la función que describe la relación entre las variables, detectando cuáles de las variables explicativas están relacionadas con la respuesta y explorando la forma e intensidad de esa relación, o bien, una vez conocida esta relación, con un fin predictivo, para conocer el valor probable de la respuesta a partir del valor conocido de los predictores. Todo análisis de regresión se concentra en el estudio de la distribución condicional de la variable respuesta a valores fijos de los predictores.

Son numerosas las aplicaciones de la regresión, y las hay en casi cualquier campo, incluyendo ingeniería, ciencias físicas y químicas, economía, administración, ciencias biológicas y ciencias sociales. De hecho, puede ser que el análisis de regresión sea la técnica estadística más usada (Montgomery et al., 2012).

La regresión lineal fue el primer tipo de análisis de regresión en ser estudiado con rigurosidad y utilizado ampliamente en aplicaciones prácticas (Yan and Su, 2009). En este enfoque, las relaciones se modelan usando funciones lineales en los parámetros. Por lo general, estos

modelos se ajustan usando el método de estimación denominado mínimos cuadrados. Las estimaciones obtenidas con este método minimizan la suma de cuadrados de los residuos de todas las observaciones, los cuales se definen como la diferencia entre la respuesta observada y la respuesta ajustada por el modelo.

El método de mínimos cuadrados surgió de los campos de la astronomía y la geodesia cuando científicos y matemáticos intentaban proporcionar soluciones a los desafíos de navegar los océanos de la Tierra durante la llamada *era del descubrimiento*, la cual tuvo lugar en Europa desde finales del siglo XV hasta finales del siglo XVIII y se caracterizó por una extensa exploración de ultramar. En esa época, la descripción precisa del comportamiento de los cuerpos celestes era la clave para permitir a los buques navegar en mar abierto, donde los marineros ya no podían confiar en avistamientos de tierra para la navegación. La primera exposición clara y concisa del método de mínimos cuadrados fue publicada por Legendre en 1805 (Legendre, 1805). En su publicación, la técnica se describe como un procedimiento algebraico para ajustar ecuaciones lineales a los datos. Legendre utilizó el método para determinar, a partir de observaciones astronómicas, la órbita de cuerpos celestes alrededor del Sol. El valor del método de mínimos cuadrados presentado por Legendre fue reconocido inmediatamente por los principales astrónomos y geodestas de la época.

En 1809, Carl Friedrich Gauss publicó su propio método para calcular la órbita de cuerpos celestes (Gauss, 1809). En su trabajo afirmó haber descubierto el método de mínimos cuadrados en 1795, lo que condujo naturalmente a una disputa de prioridad con Legendre. En la historia de la estadística, este desacuerdo recibe el nombre de "disputa de prioridad sobre el descubrimiento del método de mínimos cuadrados" (Stigler, 1981). El desarrollo de Gauss sobre este método fue superior al de Legendre, logrando su conexión con los principios de probabilidad y distribución normal.

En 1810, después de leer el trabajo de Gauss, Laplace utilizó el teorema central del límite

para justificar el uso del método de mínimos cuadrados aún cuando los errores no seguían una distribución normal (Laplace, 1810). En 1822, Gauss pudo afirmar que el enfoque mínimo-cuadrático para el ajuste de modelos lineales era óptimo en el sentido que en un modelo lineal donde los errores tienen media cero, no están correlacionados y tienen variancia constante, el mejor estimador lineal insesgado de sus coeficientes es el estimador mínimo-cuadrático. A este resultado se lo conoce como el teorema de Gauss-Markov.

La estimación mínimo-cuadrática fue la primer forma de "regresión", término introducido por Galton a finales del siglo XIX. Sir Francis Galton fue un naturalista, antropólogo, astrónomo y estadístico autodidacta. Al estudiar la altura relativa de padres e hijos, Galton observó que aquellos hijos cuyos padres tenían una estatura muy superior al promedio tendían a ser altos, pero con estaturas más cercanas al valor medio. Para aquellos hijos cuyos padres tenían una estatura muy inferior al promedio ocurría algo similar, es decir, también eran bajos pero tendían a reducir su diferencia respecto a la estatura media. A este fenómeno Galton lo llamó "regresión hacia la mediocridad", lo que en términos modernos se conoce como "regresión a la media" (Galton, 1886).

El método de mínimos cuadrados es una técnica tan popular que usualmente cuando las personas hablan de regresión lineal, en realidad se están refiriendo a la regresión mínimo-cuadrática. Esta popularidad se debe a que su aplicación es fácil de entender, provee estimaciones de parámetros que son fácilmente interpretables y su implementación en computadoras es sencilla, pudiendo resolver rápidamente problemas con cientos de predictores y cientos de miles de observaciones.

Sin embargo, existen dos motivos por los cuales el analista de los datos puede no estar satisfecho con los estimadores mínimo-cuadráticos. Uno de los ellos está relacionado con la precisión en la predicción. Los estimadores mínimo-cuadráticos no tienen sesgo pero su variancia suele ser grande en comparación con otros estimadores que no necesariamente son

insesgados (Tibshirani, 1996). La precisión en la predicción puede mejorarse reduciendo o fijando en cero alguno de los coeficientes. Si bien esto trae aparejado un incremento en el sesgo, también reduce la variancia de los valores predichos, y esta compensación mejora la precisión de los resultados obtenidos. El otro motivo tiene que ver con la interpretación y el uso de los modelos. Cuando se tiene una gran cantidad de predictores, generalmente se desea determinar un subconjunto más pequeño de variables que estén muy relacionadas con la respuesta. Al aplicar mínimos cuadrados, puede ser que un gran número de predictores resulten significativos, siendo débil su relación con la respuesta. Esto puede deberse a la existencia de predictores altamente correlacionados, fenómeno que recibe el nombre de multicolinealidad. Cuando un modelo tiene muchos predictores, es más difícil de interpretar y generalmente no resulta bueno para predecir la respuesta en nuevos vectores de variables explicativas (Hastie et al., 2009).

A través de los años, diferentes técnicas fueron desarrolladas para mejorar la estimación mínimo-cuadrática, entre ellas: selección de un subconjunto de variables, regresiones *Ridge* y LASSO (*Least Absolute Shrinkage and Selection Operator*).

Beale et al. (1967) y Hocking and Leslie (1967) fueron los primeros autores que publicaron trabajos relacionados con los procedimientos de selección de un subconjunto de variables. Este método provee modelos interpretables, pero extremadamente inestables debido a que se trata de un proceso discreto, es decir, los predictores se retienen o se excluyen del modelo, y pequeños cambios en los datos pueden resultar en la selección de modelos muy diferentes, lo que reduce la precisión de las predicciones.

La regresión *Ridge* fue presentada por Hoerl and Kennard (1970) como una alternativa a los estimadores mínimo-cuadráticos en presencia de multicolinealidad. Se trata de un proceso continuo que contrae los coeficientes, es decir, reduce su magnitud en valor absoluto, y por lo tanto es más estable. Sin embargo, no fija los coeficientes de variables muy poco asocia-

das con la respuesta exactamente en cero, razón por la cual no provee modelos fácilmente interpretables en presencia de muchas variables explicativas (Tibshirani, 1996).

En 1996, Tibshirani, intentando retener lo mejor de la selección de un subconjunto de variables y de la regresión *Ridge*, propuso la técnica denominada LASSO, la cual contrae algunos coeficientes y fija en cero a otros (Tibshirani, 1996).

En la actualidad, los grandes avances tecnológicos y la capacidad de almacenamiento creciente de los medios informáticos permite disponer de grandes bases de datos que hacen más compleja la tarea de extraer información en forma comprensible para interpretar los fenómenos investigados (Nisbet et al., 2009) (Han et al., 2011) (Leskovec et al., 2014) (Larose and Larose, 2015). En este contexto, es común encontrarse con situaciones en las que el número de variables explicativas es mucho mayor que el número de observaciones. El análisis de regresión en este escenario recibe el nombre de regresión en "grandes dimensiones".

El método de mínimos cuadrados falla en "grandes dimensiones", porque los estimadores que se obtienen no son únicos. Esta falta de unicidad de los estimadores hace que la interpretación de las soluciones pierda sentido, ya que para una solución el coeficiente estimado para un predictor puede ser positivo, mientras que para otra, puede ser negativo, es decir, el efecto de ese predictor sobre la respuesta depende de la solución elegida (Hastie et al., 2009).

Los distintos métodos de selección de un subconjunto de variables (selección del mejor subconjunto, selección paso a paso hacia adelante o hacia atrás) son impracticables en el contexto de grandes dimensiones, debido a la gran cantidad de cálculos que requieren. Por este motivo, cuando el número de variables explicativas es mayor que el número de observaciones, usualmente se recurre a las regresiones *Ridge* y LASSO.

La presente tesina está orientada al estudio de métodos de estimación en modelos de regresión que se adecúen al contexto de grandes dimensiones de datos, en particular al estudio de las regresiones *Ridge* y LASSO, haciendo estudios comparativos por simulación que

evidencien sus propiedades en cuanto a bondad de predicción del modelo y características distribucionales de los estimadores de los parámetros. También se aborda el tema de implementación de los algoritmos de estimación en la regresion LASSO, caso en el cual no existen soluciones explícitas. En el último apartado del Capítulo de Metodología se describen algunos algoritmos, así como rutinas de R disponibles. Se agregan Anexos con rutinas específicas tanto para la obtención de los estimadores, como para la implementación de las simulaciones realizadas para el estudio comparativo de propiedades.

Capítulo 2

Objetivos

2.1. Objetivo General

Estudiar distintas propuestas metodológicas para estimar los parámetros de modelos de regresión en contextos de bases de datos de grandes dimensiones donde el número de variables explicativas exceda al número de observaciones.

2.2. Objetivos Específicos

- Realizar una búsqueda bibliográfica actualizada sobre métodos de estimación en regresión adecuados para "grandes dimensiones", enfocando la atención en las regresiones *Ridge* y LASSO.
- Hacer una síntesis de los métodos y su implementación.
- Indagar sobre los programas disponibles para la obtención de los estimadores de los métodos estudiados.
- Realizar estudios por simulación que permitan verificar empíricamente las propiedades de los métodos en cuanto a su capacidad de estimar correctamente parámetros

predeterminados de un modelo.

- Estudiar propiedades distribucionales de los estimadores a través de simulaciones que contemplen distintos escenarios que representan posibles situaciones reales.

Capítulo 3

Metodología

3.1. Consideraciones Básicas

Se considera una muestra aleatoria $(\boldsymbol{x}_i^T, y_i) \in \mathbb{R}^p \times \mathbb{R}$, $i = 1, ..., n$ (Tibshirani, 2017), donde $\boldsymbol{x}_i \in \mathbb{R}^p$ es un vector columna de p variables explicativas cuantitativas continuas[1] centradas del elemento i e $y_i \in \mathbb{R}$ es una variable respuesta, también continua, de ese mismo elemento. Las variables $\boldsymbol{x}_i$ e y_i se suponen relacionadas por el modelo

$$y_i = \boldsymbol{x}_i^T \boldsymbol{\beta} + \epsilon_i \qquad i = 1, ..., n \tag{3.1}$$

donde $\boldsymbol{\beta} \in \mathbb{R}^p$ es un vector desconocido de coeficientes y ϵ_i, $i = 1, ..., n$, son errores aleatorios con $E(\epsilon_i) = 0$. Sin pérdida de generalidad se ignora el término correspondiente a la ordenada al origen. Escrito en forma matricial el modelo resulta

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{3.2}$$

[1] Si bien no es necesario que las variables explicativas sean continuas, en este trabajo sólo se considera este caso.

donde $\boldsymbol{y} = (y_1, ..., y_n)^T \in \mathbb{R}^n$ es el vector de variables respuestas, $\boldsymbol{X} \in \mathbb{R}^{n \times p}$ es la matriz de variables explicativas centradas, con i-ésima fila $\boldsymbol{x}_i^T \in \mathbb{R}^p$, $i = 1, ..., n$, y j-ésima columna $\boldsymbol{X}_j \in \mathbb{R}^n$, $j = 1, ..., p$, y $\boldsymbol{\epsilon} = (\epsilon_1, ..., \epsilon_n)^T \in \mathbb{R}^n$ es el vector de errores aleatorios.

Usualmente se supone que los valores de las variables $\boldsymbol{x}_i$, $i = 1, ..., n$, son predeterminados y que los errores ϵ_i, $i = 1, ..., n$, son independientes e igualmente distribuidos ($i.i.d.$). Por lo general, se asume que la distribución de los errores es Normal con variancia constante σ^2.

3.2. Regresión Mínimo Cuadrática

Considerando predeterminada a la matriz de predictores $\boldsymbol{X}$, los estimadores mínimo-cuadráticos de los coeficientes del modelo (3.1) son aquellos coeficientes $\boldsymbol{\beta}$ que minimizan la Suma de Cuadrados Residual (SCR) a través de todas las observaciones, es decir, se definen como la solución al siguiente problema de optimización

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^p} SCR(\boldsymbol{\beta}) \iff \min_{\boldsymbol{\beta} \in \mathbb{R}^p} \sum_{i=1}^{n} (y_i - \boldsymbol{x}_i^T \boldsymbol{\beta})^2 \iff \min_{\boldsymbol{\beta} \in \mathbb{R}^p} ||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}||_2^2, \tag{3.3}$$

donde $||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}||_2 = \sqrt{\sum_{i=1}^{n} (y_i - \boldsymbol{x}_i^T \boldsymbol{\beta})^2}$ es la denominada norma 2 de $\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}$.

Si el rango de la matriz $\boldsymbol{X}$ es igual a p, es decir, si los predictores $\boldsymbol{X}_1$, ..., $\boldsymbol{X}_p$ (columnas de la matriz $\boldsymbol{X}$, cada una de dimensión $n \times 1$) son linealmente independientes, entonces el problema de optimización mínimo-cuadrático tiene solución única

$$\hat{\boldsymbol{\beta}}^{MC} = (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{y}. \tag{3.4}$$

El estimador $\hat{\boldsymbol{\beta}}^{MC}$ es insesgado de $\boldsymbol{\beta}$, ya que

$$E\left(\hat{\boldsymbol{\beta}}^{MC} | \boldsymbol{X}\right) = E\left((\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{y} | \boldsymbol{X}\right) = (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T E(\boldsymbol{y} | \boldsymbol{X}) =$$

$$= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{\beta}. \tag{3.5}$$

Suponiendo que $\boldsymbol{\epsilon} \sim N(\boldsymbol{0}, \sigma^2\boldsymbol{I})$, la matriz de variancias y covariancias de $\hat{\boldsymbol{\beta}}^{MC}$ resulta

$$Cov\left(\hat{\boldsymbol{\beta}}^{MC}|\boldsymbol{X}\right) = Cov\left((\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}|\boldsymbol{X}\right) = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\, Cov(\boldsymbol{y}|\boldsymbol{X})\, \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1} =$$

$$= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\, \sigma^2\, \boldsymbol{I}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1} = \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}. \tag{3.6}$$

Usando la Descomposición en Valores Singulares (DVS) de una matriz $\boldsymbol{X}$ de dimensión $n \times p$, la cual tiene la forma

$$\boldsymbol{X} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^T, \tag{3.7}$$

donde $\boldsymbol{U}$ es una matriz de dimensión $n \times p$ cuyas columnas son ortonormales, $\boldsymbol{V}$ es una matriz ortogonal de dimensión $p \times p$ y $\boldsymbol{D}$ es una matriz diagonal de dimensión $p \times p$ que contiene los valores singulares de $\boldsymbol{X}$, la ecuación (3.6) resulta

$$\sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1} = \sigma^2(\boldsymbol{V}\boldsymbol{D}\boldsymbol{U}^T\boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^T)^{-1} = \sigma^2(\boldsymbol{V}\boldsymbol{D}\boldsymbol{I}\boldsymbol{D}\boldsymbol{V}^T)^{-1} = \sigma^2(\boldsymbol{V}\boldsymbol{D}^2\boldsymbol{V}^T)^{-1}. \tag{3.8}$$

La Variancia Total de $\hat{\boldsymbol{\beta}}^{MC}$, definida como la traza de $Cov\left(\hat{\boldsymbol{\beta}}^{MC}|\boldsymbol{X}\right)$, resulta

$$\text{tr}\left(Cov\left(\hat{\boldsymbol{\beta}}^{MC}|\boldsymbol{X}\right)\right) = \text{tr}\left(\sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}\right) = \text{tr}\left(\sigma^2(\boldsymbol{V}\boldsymbol{D}^2\boldsymbol{V}^T)^{-1}\right) = \sigma^2\sum_{j=1}^{p}\frac{1}{d_j^2}, \tag{3.9}$$

donde d_j es el j-ésimo elemento diagonal de $\boldsymbol{D}$.

Los valores predichos a través del modelo (3.1), también llamados valores ajustados, son $\hat{\boldsymbol{y}}^{MC} = \boldsymbol{X}\hat{\boldsymbol{\beta}}^{MC} = \boldsymbol{H}\,\boldsymbol{y}$, donde $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$ es la matriz de proyección sobre el espacio de las columnas de $\boldsymbol{X}$. Estos valores ajustados son las predicciones en los puntos de la muestra $\boldsymbol{x}_i$, $i = 1, ..., n$. La predicción sobre una nueva observación $\boldsymbol{x}_0 \in \mathbb{R}^p$ resulta

$\boldsymbol{x}_0^T\hat{\boldsymbol{\beta}}^{MC}$.

Estas predicciones son insesgadas, dado que:

$$E(\boldsymbol{x}^T\hat{\boldsymbol{\beta}}^{MC}|\boldsymbol{X}) = E(\boldsymbol{x}^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}|\boldsymbol{X}) = \boldsymbol{x}^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^TE(\boldsymbol{y}|\boldsymbol{X}) =$$

$$= \boldsymbol{x}^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{x}^T\boldsymbol{\beta}, \tag{3.10}$$

para cualquier vector de predictores $\boldsymbol{x} \in \mathbb{R}^p$. Este insesgamiento no requiere suponer independencia entre $\boldsymbol{X}$ y $\boldsymbol{\epsilon}$.

El error de predicción para las observaciones de la muestra, también denominado error de entrenamiento del estimador mínimo-cuadrático se define como

$$\frac{1}{n}E||\hat{\boldsymbol{y}}^{MC} - \boldsymbol{y}||_2^2 = \frac{1}{n}E||\boldsymbol{X}\hat{\boldsymbol{\beta}}^{MC} - \boldsymbol{X}\boldsymbol{\beta}||_2^2, \tag{3.11}$$

donde $\boldsymbol{X}$ es considerada aleatoria e independiente de $\boldsymbol{\epsilon}$.

Condicionando a $\boldsymbol{X}$ en (3.11), para mantener fijos los valores de esta matriz, resulta

$$\frac{1}{n}E(||\boldsymbol{X}\hat{\boldsymbol{\beta}}^{MC} - \boldsymbol{X}\boldsymbol{\beta}||_2^2|\ \boldsymbol{X}) = \frac{1}{n}E(||\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}||_2^2|\ \boldsymbol{X}), \tag{3.12}$$

lo cual puede escribirse como

$$\frac{1}{n}E\left(\sum_{i=1}^{n}(\boldsymbol{x}_i^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y} - \boldsymbol{x}_i^T\boldsymbol{\beta})^2|\ \boldsymbol{X}\right) = \frac{1}{n}\sum_{i=1}^{n}\left(E(\boldsymbol{x}_i^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y} - \boldsymbol{x}_i^T\boldsymbol{\beta})^2|\ \boldsymbol{X}\right) =$$

$$= \frac{1}{n}\sum_{i=1}^{n}\left(Var(\boldsymbol{x}_i^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y})|\boldsymbol{X}\right). \tag{3.13}$$

Definiendo con $Cov(\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}|\boldsymbol{X})$ a la matriz de variancias y covariancias de

$\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$ condicionada a $\boldsymbol{X}$ y por definición de $\boldsymbol{H}$, la ecuación (3.13) resulta

$$\frac{1}{n}\text{tr}\left(Cov(\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}|\boldsymbol{X})\right) = \frac{1}{n}\text{tr}\left(Cov(\boldsymbol{H}\,\boldsymbol{y}|\boldsymbol{X})\right) = \frac{1}{n}\text{tr}\left(\boldsymbol{H}\,Cov(\boldsymbol{y})\,\boldsymbol{H}|\boldsymbol{X}\right) =$$
$$= \frac{1}{n}\text{tr}\left(\boldsymbol{H}\,Cov(\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{\epsilon})\,\boldsymbol{H}|\boldsymbol{X}\right) = \frac{1}{n}\text{tr}\left(\boldsymbol{H}\,Cov(\boldsymbol{\epsilon})\,\boldsymbol{H}\right) = \frac{1}{n}\text{tr}\left(\boldsymbol{H}\,\sigma^2\boldsymbol{I}\,\boldsymbol{H}\right) =$$
$$= \sigma^2\frac{1}{n}\text{tr}\left(\boldsymbol{H}\,\boldsymbol{H}\right) = \sigma^2\frac{1}{n}\text{tr}\left(\boldsymbol{H}\right) = \sigma^2\frac{1}{n}\text{tr}\left(\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\right) =$$
$$= \sigma^2\frac{1}{n}\text{tr}\left(\boldsymbol{X}^T\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\right) = \sigma^2\frac{1}{n}\text{tr}\left(\boldsymbol{I}_p\right) = \sigma^2\frac{p}{n} \tag{3.14}$$

donde $\boldsymbol{I}_p$ es la matriz identidad de dimensión p.

El resultado obtenido en (3.14) no depende de $\boldsymbol{X}$, por lo tanto, el error de predicción del estimador mínimo-cuadrático para las observaciones de la muestra resulta igual a

$$\frac{1}{n}E||\boldsymbol{X}\hat{\boldsymbol{\beta}}^{MC} - \boldsymbol{X}\boldsymbol{\beta}||_2^2 = \sigma^2\frac{p}{n}. \tag{3.15}$$

El error de predicción para vectores de variables explicativas no observados, también denominado riesgo predictivo del estimador mínimo-cuadrático se define como

$$E(\boldsymbol{x}_0^T\hat{\boldsymbol{\beta}}^{MC} - \boldsymbol{x}_0^T\boldsymbol{\beta})^2, \tag{3.16}$$

donde $\boldsymbol{x}_0 \in \mathbb{R}^p$ es una nueva observación del conjunto de variables explicativas.

Condicionando a $\boldsymbol{X}$ y $\boldsymbol{x}_0$ en (3.16) se obtiene que

$$E(\boldsymbol{x}_0^T\hat{\boldsymbol{\beta}}^{MC} - \boldsymbol{x}_0^T\boldsymbol{\beta}|\boldsymbol{X},\boldsymbol{x}_0)^2 = Var(\boldsymbol{x}_0^T\hat{\boldsymbol{\beta}}^{MC}|\boldsymbol{X},\boldsymbol{x}_0) = Var\left(\boldsymbol{x}_0^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}|\boldsymbol{X},\boldsymbol{x}_0\right) =$$
$$= \boldsymbol{x}_0^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\,Var(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{x}_0)\,\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_0 = \boldsymbol{x}_0^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\,\sigma^2\,\boldsymbol{I}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_0 =$$
$$= \sigma^2\boldsymbol{x}_0^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_0 = \sigma^2\boldsymbol{x}_0^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_0. \tag{3.17}$$

El resultado obtenido en (3.17) es un escalar, por lo que $\boldsymbol{x}_0^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_0$ puede escribir-

se como $\text{tr}(\boldsymbol{x}_0^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_0)$. La matriz de predictores $\boldsymbol{X}$ usada en la estimación mínimo-cuadrática y el vector de variables $\boldsymbol{x}_0$ correspondiente a la nueva observación del conjunto de variables explicativas son independientes.

Luego, al integrar sobre $\boldsymbol{X}$ y $\boldsymbol{x}_0$ en (3.17), por propiedades de traza de una matriz y de esperanza, resulta

$$E(\boldsymbol{x}_0^T\hat{\boldsymbol{\beta}}^{MC} - \boldsymbol{x}_0^T\boldsymbol{\beta})^2 = \sigma^2 E[\text{tr}(\boldsymbol{x}_0^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_0)] = \sigma^2 E[\text{tr}(\boldsymbol{x}_0\boldsymbol{x}_0^T(\boldsymbol{X}^T\boldsymbol{X})^{-1})] =$$
$$= \sigma^2 \text{tr}\left(E(\boldsymbol{x}_0\boldsymbol{x}_0^T)E[(\boldsymbol{X}^T\boldsymbol{X})^{-1}]\right). \tag{3.18}$$

A diferencia de lo observado con el error de predicción dentro de la muestra, el riesgo predictivo depende de la distribución de los predictores, por lo que no es posible obtener una fórmula general exacta.

Sea $\boldsymbol{\Sigma}$ la matriz de covariancia de los predictores. Como se asume que los predictores están centrados, $\boldsymbol{\Sigma} = E(\boldsymbol{x}_0\boldsymbol{x}_0^T) = \frac{1}{n}E(\boldsymbol{X}^T\boldsymbol{X})$. Por lo tanto, el error de predicción fuera de la muestra en la ecuación (3.18) resulta igual a

$$\sigma^2 \text{tr}\left(\boldsymbol{\Sigma}\, E[(\boldsymbol{X}^T\boldsymbol{X})^{-1}]\right). \tag{3.19}$$

Groves and Rothenberg (1969) demostraron que la matriz que se obtiene al hacer la diferencia $E[(\boldsymbol{X}^T\boldsymbol{X})^{-1}] - [E(\boldsymbol{X}^T\boldsymbol{X})]^{-1}$ es semidefinida positiva. Usando este resultado y el de la ecuación (3.19) resulta

$$E(\boldsymbol{x}_0^T\hat{\boldsymbol{\beta}}^{MC} - \boldsymbol{x}_0^T\boldsymbol{\beta})^2 = \sigma^2 \text{tr}\left(\boldsymbol{\Sigma}\, E[(\boldsymbol{X}^T\boldsymbol{X})^{-1}]\right) \geq \sigma^2 \text{tr}\left(\boldsymbol{\Sigma}\,[E(\boldsymbol{X}^T\boldsymbol{X})]^{-1}\right) =$$
$$= \sigma^2 \text{tr}\left(\boldsymbol{\Sigma}\,[n\,\boldsymbol{\Sigma}]^{-1}\right) = \sigma^2 \frac{1}{n}\text{tr}\left(\boldsymbol{\Sigma}\,\boldsymbol{\Sigma}^{-1}\right) = \sigma^2\frac{p}{n}. \tag{3.20}$$

Así queda demostrado que el riesgo predictivo es siempre mayor o igual que el error de

predicción para las observaciones de la muestra, lo cual tiene sentido ya que intuitivamente realizar predicciones sobre nuevas observaciones es menos preciso que obtener los valores ajustados para los mismos valores que fueron utilizados en la estimación del modelo.

Si los predictores son variables aleatorias $N(\mathbf{0}, \boldsymbol{\Sigma})$ independientes de ϵ_i, el error de predicción para vectores de variables explicativas no observados puede calcularse de forma exacta. Resulta que $\boldsymbol{X}^T\boldsymbol{X} \sim W(\boldsymbol{\Sigma}, n)$, siendo W una distribución Wishart (Wishart, 1928), y $(\boldsymbol{X}^T\boldsymbol{X})^{-1} \sim W^{-1}(\boldsymbol{\Sigma}^{-1}, n)$, siendo W^{-1} una distribución Wishart inversa (Haff, 1979), por lo que

$$E(\boldsymbol{x}_0^T\hat{\boldsymbol{\beta}}^{MC} - \boldsymbol{x}_0^T\boldsymbol{\beta})^2 = \sigma^2 \text{tr}\left(\boldsymbol{\Sigma}\frac{\boldsymbol{\Sigma}^{-1}}{n-p-1}\right) = \sigma^2\frac{p}{n-p-1}. \tag{3.21}$$

3.3. Problemas del Ajuste Mínimo Cuadrático en Grandes Dimensiones

Cuando el rango de la matriz $\boldsymbol{X}$ es menor que p, hecho que ocurre cuando $p > n$, existen infinitas soluciones al problema de optimización mínimo-cuadrático (3.3). Dada una solución $\hat{\boldsymbol{\beta}}$, $\hat{\boldsymbol{\beta}} + \boldsymbol{\eta}$ también es solución para cualquier $\boldsymbol{\eta} \in null(\boldsymbol{X}) = \{\boldsymbol{\eta} \in \mathbb{R}^p : \boldsymbol{X}\boldsymbol{\eta} = \mathbf{0}\}$, donde $null(\boldsymbol{X})$ es el espacio nulo de la matriz $\boldsymbol{X}$, ya que

$$\boldsymbol{X}(\hat{\boldsymbol{\beta}} + \boldsymbol{\eta}) = \boldsymbol{X}\hat{\boldsymbol{\beta}} + \boldsymbol{X}\boldsymbol{\eta} = \boldsymbol{X}\hat{\boldsymbol{\beta}}, \tag{3.22}$$

es decir, con $\hat{\boldsymbol{\beta}}$ y $\hat{\boldsymbol{\beta}}+\boldsymbol{\eta}$ se alcanza el mismo mínimo en (3.3). Al no haber una solución única, la interpretación de las soluciones pierde sentido, ya que para al menos un $j \in \{1, ..., p\}$ se tendrá que $\hat{\beta}_j > 0$ para una solución $\hat{\boldsymbol{\beta}}$, pero $\tilde{\beta}_j < 0$ para otra solución $\tilde{\boldsymbol{\beta}}$, es decir, una variable explicativa influiría en forma directa e inversa sobre la respuesta simultáneamente.

A pesar de esta falta de unicidad de los estimadores de $\boldsymbol{\beta}$, los valores ajustados por regresión mínimo-cuadrática son siempre únicos en los puntos de la muestra, es decir, $\boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}\tilde{\boldsymbol{\beta}}$

para dos soluciones cualesquiera $\hat{\boldsymbol{\beta}}$ y $\tilde{\boldsymbol{\beta}}$, sin importar el rango de las columnas de $\boldsymbol{X}$. Esto es así porque los valores ajustados siempre pueden escribirse como $\boldsymbol{H}\,\boldsymbol{y}$, donde $\boldsymbol{H}$ es la matriz de proyección sobre el espacio de las columnas de $\boldsymbol{X}$. Esta matriz resulta igual a $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^-\boldsymbol{X}^T$, donde $(\boldsymbol{X}^T\boldsymbol{X})^-$ es la inversa generalizada de $\boldsymbol{X}^T\boldsymbol{X}$. Cuando $\boldsymbol{X}$ es de rango completo en las columnas puede escribirse $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$.

Sin embargo, en términos de predicciones propiamente dichas, es decir, para una nueva observación $\boldsymbol{x}_0 \in \mathbb{R}^p$ no perteneciente a las filas de $\boldsymbol{X}$, generalmente no será cierto que $\boldsymbol{x}_0\hat{\boldsymbol{\beta}} = \boldsymbol{x}_0\tilde{\boldsymbol{\beta}}$ para dos soluciones $\hat{\boldsymbol{\beta}}$ y $\tilde{\boldsymbol{\beta}}$ debido a que esas soluciones $\hat{\boldsymbol{\beta}}$ y $\tilde{\boldsymbol{\beta}}$ no son necesariamente iguales.

De este modo, tanto la interpretación de los resultados como la realización de predicciones sobre nuevas observaciones pierden sentido al usar mínimos cuadrados cuando $p > n$, lo cual es un problema importante.

Incluso cuando el rango de la matriz $\boldsymbol{X}$ es igual a p, situación en la cual existe una solución única al problema de optimización mínimo-cuadrático, puede que no sea conveniente utilizar mínimos cuadrados si p es muy cercano a n, debido a que el error de predicción para las observaciones de la muestra puede ser alto. Se recuerda que este error es igual a $\sigma^2\dfrac{p}{n}$ (ec. 3.15), el cual es grande si p se aproxima demasiado a n.

Para tratar estos problemas, se pueden utilizar métodos de regresión penalizada, también denominados métodos de regularización. Al utilizar estos métodos, el estimador mínimo-cuadrático se modifica en una de dos formas:

- Forma restringida:

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^p} ||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}||_2^2 \quad \text{sujeto a} \quad \boldsymbol{\beta} \in C, \tag{3.23}$$

donde C es algún conjunto usualmente convexo.

- Forma penalizada:

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^p} ||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}||_2^2 + P(\boldsymbol{\beta}), \tag{3.24}$$

donde $P(\cdot)$ es alguna función de penalización usualmente convexa.

Los métodos de regularización buscan una reducción importante en la variancia de las estimaciones a costa de la introducción de un sesgo de moderada magnitud, lo que mejora los resultados globalmente.

En las secciones siguientes se trabaja con dos métodos particulares de regularización denominados regresión *Ridge* y regresión LASSO.

3.4. Regresión en Grandes Dimensiones

Por lo general, en términos de regularización, el conjunto C en (3.23) se elige imponiendo una restricción (cota superior) a una norma y la función de penalización $P(\cdot)$ en (3.24) se elige de modo que P sea el múltiplo de una norma.

Sean las normas

$$||\boldsymbol{\beta}||_0 = \sum_{j=1}^{p} I\{\beta_j \neq 0\}, \quad ||\boldsymbol{\beta}||_1 = \sum_{j=1}^{p} |\beta_j|, \quad ||\boldsymbol{\beta}||_2 = \left(\sum_{j=1}^{p} \beta_j^2\right)^{1/2}$$

denominadas ℓ_0, ℓ_1 y ℓ_2 respectivamente. La función que define ℓ_0 no es una norma propiamente dicha, ya que no satisface la propiedad de homogeneidad positiva, es decir, $||a\boldsymbol{\beta}||_0 \neq a||\boldsymbol{\beta}||_0$ para a distinto de 0 o 1.

En forma restringida, las normas definidas dan origen a los siguientes problemas de optimización, donde r y t son constantes tales que $r \in \mathbb{N}_0$ y $t \in \mathbb{R}_0^+$:

- Selección del mejor subconjunto:

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^p} ||\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}||_2^2 \quad \text{sujeto a} \quad ||\boldsymbol{\beta}||_0 \leq r \tag{3.25}$$

- Regresión LASSO:

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^p} ||\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}||_2^2 \quad \text{sujeto a} \quad ||\boldsymbol{\beta}||_1 \leq t \tag{3.26}$$

- Regresión *Ridge*:

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^p} ||\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}||_2^2 \quad \text{sujeto a} \quad ||\boldsymbol{\beta}||_2^2 \leq t \tag{3.27}$$

Las constantes r y t son llamadas parámetros de suavizado. En (3.25) tiene sentido restringir r de modo que sea un número entero, ya que en la selección del mejor subconjunto se está buscando el mejor subconjunto de variables de tamaño r, en términos del error de entrenamiento obtenido.

En forma penalizada, el uso de las normas ℓ_0, ℓ_1 y ℓ_2 da origen a los siguientes problemas de optimización, donde $\lambda \geq 0$ es una constante denominada parámetro de suavizado:

- Selección del mejor subconjunto:

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \frac{1}{2}||\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}||_2^2 + \lambda||\boldsymbol{\beta}||_0 \tag{3.28}$$

- Regresión LASSO:

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \frac{1}{2}||\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}||_2^2 + \lambda||\boldsymbol{\beta}||_1 \tag{3.29}$$

- Regresión *Ridge*:

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \frac{1}{2}||\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}||_2^2 + \lambda||\boldsymbol{\beta}||_2^2 \tag{3.30}$$

Los problemas (3.26) y (3.27) son equivalentes a los problemas (3.29) y (3.30) respec-

tivamente (Kloft et al., 2009). Esto quiere decir que para cualquier $t \geq 0$ y solución $\hat{\boldsymbol{\beta}}$ en (3.26) existe un valor $\lambda \geq 0$ tal que $\hat{\boldsymbol{\beta}}$ también es solución de (3.29), y viceversa. La misma relación de equivalencia vale para (3.27) y (3.30).

Los factores $\frac{1}{2}$ que multiplican la SCR en los problemas (3.29) y (3.30) no influyen en los resultados, solamente se los utiliza para simplificar el procedimiento de obtención de los extremos.

Los problemas (3.25) y (3.28) no son equivalentes. Para todo valor $\lambda \geq 0$ y solución $\hat{\boldsymbol{\beta}}$ en (3.28) existe un valor de $r \geq 0$ tal que $\hat{\boldsymbol{\beta}}$ también es solución de (3.25), pero no vale la inversa. Esto se debe a que el problema original no es convexo (Hastie et al., 2015).

3.4.1. Regresión *Ridge*

El estimador *Ridge* se define como

$$\hat{\boldsymbol{\beta}}^{Ridge} = \underset{\boldsymbol{\beta} \in \mathbb{R}^p}{\operatorname{argmin}} \left\{ \frac{1}{2} \sum_{i=1}^{n} (y_i - \boldsymbol{x}_i^T \boldsymbol{\beta})^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\}. \tag{3.31}$$

El criterio a minimizar en la ecuación (3.31) escrito en forma matricial resulta igual a

$$\frac{1}{2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) + \lambda \boldsymbol{\beta}^T \boldsymbol{\beta}. \tag{3.32}$$

Derivando con respecto a $\boldsymbol{\beta}$ e igualando a cero se obtiene

$$(\boldsymbol{X}^T \boldsymbol{X})\boldsymbol{\beta} - \boldsymbol{X}^T \boldsymbol{y} + 2\lambda \boldsymbol{\beta} = 0, \tag{3.33}$$

o bien:

$$(\boldsymbol{X}^T \boldsymbol{X})\boldsymbol{\beta} + 2\lambda \boldsymbol{\beta} = \boldsymbol{X}^T \boldsymbol{y} \tag{3.34}$$

$$(\boldsymbol{X}^T\boldsymbol{X} + 2\lambda\boldsymbol{I})\boldsymbol{\beta} = \boldsymbol{X}^T\boldsymbol{y}. \tag{3.35}$$

Por lo tanto, el estimador *Ridge* resulta

$$\hat{\boldsymbol{\beta}}^{Ridge} = (\boldsymbol{X}^T\boldsymbol{X} + 2\lambda\boldsymbol{I})^{-1}\boldsymbol{X}^T\boldsymbol{y}, \tag{3.36}$$

el cual es una función lineal de $\boldsymbol{y}$. El estimador $\hat{\boldsymbol{\beta}}^{Ridge}$ se acerca a $\hat{\boldsymbol{\beta}}^{MC}$ a medida que λ se acerca a cero.

Utilizando la DVS de la matriz $\boldsymbol{X}$ (ec. 3.7), otro modo de expresar $\hat{\boldsymbol{\beta}}^{Ridge}$ es

$$\begin{aligned}
\hat{\boldsymbol{\beta}}^{Ridge} = (\boldsymbol{X}^T\boldsymbol{X} + 2\lambda\boldsymbol{I})^{-1}\boldsymbol{X}^T\boldsymbol{y} = (\boldsymbol{V}\boldsymbol{D}\boldsymbol{U}^T\boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^T + 2\lambda\boldsymbol{I})^{-1}\boldsymbol{V}\boldsymbol{D}\boldsymbol{U}^T\boldsymbol{y} = \\
= (\boldsymbol{V}\boldsymbol{D}^2\boldsymbol{V}^T + 2\lambda\boldsymbol{I})^{-1}\boldsymbol{V}\boldsymbol{D}\boldsymbol{U}^T\boldsymbol{y} = (\boldsymbol{V}\boldsymbol{D}^2\boldsymbol{V}^T + 2\lambda\boldsymbol{V}\boldsymbol{V}^T)^{-1}\boldsymbol{V}\boldsymbol{D}\boldsymbol{U}^T\boldsymbol{y} = \\
= \boldsymbol{V}(\boldsymbol{D}^2 + 2\lambda\boldsymbol{I})^{-1}\boldsymbol{V}^T\boldsymbol{V}\boldsymbol{D}\boldsymbol{U}^T\boldsymbol{y} = \boldsymbol{V}(\boldsymbol{D}^2 + 2\lambda\boldsymbol{I})^{-1}\boldsymbol{D}\boldsymbol{U}^T\boldsymbol{y} = \\
= \boldsymbol{V}\ diag\left(\frac{d_j}{d_j^2 + 2\lambda}\right)\ \boldsymbol{U}^T\boldsymbol{y}.
\end{aligned} \tag{3.37}$$

El estimador *Ridge* minimiza una SCR Penalizada

$$SCR(\boldsymbol{\beta}, \lambda) = ||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}||_2^2 + \lambda||\boldsymbol{\beta}||_2^2 = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) + \lambda\boldsymbol{\beta}^T\boldsymbol{\beta}. \tag{3.38}$$

La penalización cuadrática $\boldsymbol{\beta}^T\boldsymbol{\beta}$ hace que la solución *Ridge* dependa de la escala de los predictores, razón por la cual en (3.31) se trabaja con predictores estandarizados.

La ordenada al origen se excluye del término de penalización para evitar que el procedimiento dependa del origen elegido para $\boldsymbol{y}$. Si la ordenada al origen no se excluye del término de penalización, la suma de una constante c a cada una de las y_i no resulta simplemente en la suma de esa constante c a las predicciones obtenidas usando las y_i (Hastie et al., 2009).

Por lo general, la matriz $\boldsymbol{X}^T\boldsymbol{X}+2\lambda\boldsymbol{I}$ es invertible, incluso cuando $\boldsymbol{X}^T\boldsymbol{X}$ es singular. Esto se debe a la suma de una constante positiva a la diagonal de $\boldsymbol{X}^T\boldsymbol{X}$. Este aspecto particular fue el que motivó el uso de la regresión *Ridge* cuando se introdujo en estadística (Hoerl and Kennard, 1970).

Asumiendo que $\boldsymbol{X}^T\boldsymbol{X}$ es no singular, el estimador *Ridge* se relaciona con $\hat{\boldsymbol{\beta}}^{MC}$ del siguiente modo

$$\hat{\boldsymbol{\beta}}^{Ridge} = [\boldsymbol{I} + 2\lambda(\boldsymbol{X}^T\boldsymbol{X})^{-1}]^{-1}\hat{\boldsymbol{\beta}}^{MC}, \tag{3.39}$$

Esta relación se verifica reemplazando $\hat{\boldsymbol{\beta}}^{MC}$ por su definición y usando la propiedad de inversa de producto de matrices $\left((\boldsymbol{AB})^{-1} = \boldsymbol{B}^{-1}\boldsymbol{A}^{-1}\right)$ sobre $\boldsymbol{A} = \boldsymbol{X}^T\boldsymbol{X}$ y $\boldsymbol{B} = \boldsymbol{I} + 2\lambda(\boldsymbol{X}^T\boldsymbol{X})^{-1}$.

Cuando $\boldsymbol{X}$ es una matriz ortogonal, se verifica que $\boldsymbol{X}^T\boldsymbol{X} = \boldsymbol{I}$, y la solución mínimo cuadrática (3.4) resulta

$$\hat{\boldsymbol{\beta}}^{MC} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y} = \boldsymbol{X}^T\boldsymbol{y}. \tag{3.40}$$

Luego, el estimador *Ridge* cuando la matriz $\boldsymbol{X}$ es ortogonal resulta

$$\hat{\boldsymbol{\beta}}^{Ridge} = (\boldsymbol{I} + 2\lambda\boldsymbol{I})^{-1}\boldsymbol{X}^T\boldsymbol{y} = ((1+2\lambda)\boldsymbol{I})^{-1}\boldsymbol{X}^T\boldsymbol{y} = \frac{1}{1+2\lambda}\hat{\boldsymbol{\beta}}^{MC}. \tag{3.41}$$

Por lo tanto, en este contexto, el estimador *Ridge* es simplemente una versión ponderada del estimador mínimo cuadrático.

El estimador *Ridge* es sesgado del verdadero vector de coeficientes $\boldsymbol{\beta_0}$. Usando los supuestos del modelo (3.1) y la definición de $\hat{\boldsymbol{\beta}}^{Ridge}$ se obtiene

$$E\left(\hat{\boldsymbol{\beta}}^{Ridge}|\boldsymbol{X}\right) = E\left((\boldsymbol{X}^T\boldsymbol{X} + 2\lambda\boldsymbol{I})^{-1}\boldsymbol{X}^T\boldsymbol{y}|\boldsymbol{X}\right) = (\boldsymbol{X}^T\boldsymbol{X} + 2\lambda\boldsymbol{I})^{-1}\boldsymbol{X}^T E(\boldsymbol{y}|\boldsymbol{X}) =$$

$$= (\boldsymbol{X}^T\boldsymbol{X} + 2\lambda\boldsymbol{I})^{-1}\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta} = (\boldsymbol{X}^T\boldsymbol{X} + 2\lambda\boldsymbol{I})^{-1}(\boldsymbol{X}^T\boldsymbol{X} + 2\lambda\boldsymbol{I} - 2\lambda\boldsymbol{I})\boldsymbol{\beta} =$$

$$= \left((\boldsymbol{X}^T\boldsymbol{X} + 2\lambda\boldsymbol{I})^{-1}(\boldsymbol{X}^T\boldsymbol{X} + 2\lambda\boldsymbol{I}) - (\boldsymbol{X}^T\boldsymbol{X} + 2\lambda\boldsymbol{I})^{-1}2\lambda\boldsymbol{I}\right)\boldsymbol{\beta} =$$

$$= \left(\boldsymbol{I} - 2\lambda(\boldsymbol{X}^T\boldsymbol{X} + 2\lambda\boldsymbol{I})^{-1}\right)\boldsymbol{\beta} = \boldsymbol{\beta} - 2\lambda(\boldsymbol{X}^T\boldsymbol{X} + 2\lambda\boldsymbol{I})^{-1}\boldsymbol{\beta}. \tag{3.42}$$

El sesgo del estimador *Ridge* es proporcional a λ. Cuanto mayor sea el valor de λ, mayor será el sesgo de $\hat{\boldsymbol{\beta}}^{Ridge}$ con respecto a $\boldsymbol{\beta}$.

Suponiendo que $\boldsymbol{\epsilon} \sim N(\boldsymbol{0}, \sigma^2\boldsymbol{I})$, la matriz de variancias y covariancias de $\hat{\boldsymbol{\beta}}^{Ridge}$ resulta

$$\begin{aligned} Cov\left(\hat{\boldsymbol{\beta}}^{Ridge}|\boldsymbol{X}\right) &= Cov\left((\boldsymbol{X}^T\boldsymbol{X} + 2\lambda\boldsymbol{I})^{-1}\boldsymbol{X}^T\boldsymbol{y}|\boldsymbol{X}\right) = \\ &= (\boldsymbol{X}^T\boldsymbol{X} + 2\lambda\boldsymbol{I})^{-1}\boldsymbol{X}^T Cov(\boldsymbol{y}|\boldsymbol{X})\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X} + 2\lambda\boldsymbol{I})^{-1} = \\ &= \sigma^2(\boldsymbol{X}^T\boldsymbol{X} + 2\lambda\boldsymbol{I})^{-1}(\boldsymbol{X}^T\boldsymbol{X})(\boldsymbol{X}^T\boldsymbol{X} + 2\lambda\boldsymbol{I})^{-1}. \end{aligned} \tag{3.43}$$

Usando la DVS de la matriz $\boldsymbol{X}$ (ec. 3.7),

$$(\boldsymbol{X}^T\boldsymbol{X} + 2\lambda\boldsymbol{I})^{-1} = (\boldsymbol{V}\boldsymbol{D}^2\boldsymbol{V}^T + 2\lambda\boldsymbol{I})^{-1} = \boldsymbol{V}(\boldsymbol{D}^2 + 2\lambda\boldsymbol{I})^{-1}\boldsymbol{V}^T, \tag{3.44}$$

de modo que la ecuación (3.43) puede escribirse como

$$\sigma^2\boldsymbol{V}(\boldsymbol{D}^2 + 2\lambda\boldsymbol{I})^{-1}\boldsymbol{V}^T\boldsymbol{V}\boldsymbol{D}^2\boldsymbol{V}^T\boldsymbol{V}(\boldsymbol{D}^2 + 2\lambda\boldsymbol{I})^{-1}\boldsymbol{V}^T, \tag{3.45}$$

lo cual resulta igual a

$$\sigma^2\boldsymbol{V}(\boldsymbol{D}^2 + 2\lambda\boldsymbol{I})^{-1}\boldsymbol{D}^2(\boldsymbol{D}^2 + 2\lambda\boldsymbol{I})^{-1}\boldsymbol{V}^T. \tag{3.46}$$

La Variancia Total de $\hat{\boldsymbol{\beta}}^{Ridge}$, definida como la traza de $Cov\left(\hat{\boldsymbol{\beta}}^{Ridge}|\boldsymbol{X}\right)$, resulta

$$\begin{aligned} \text{tr}\left(Cov\left(\hat{\boldsymbol{\beta}}^{Ridge}|\boldsymbol{X}\right)\right) &= \text{tr}\left(\sigma^2\boldsymbol{V}(\boldsymbol{D}^2 + 2\lambda\boldsymbol{I})^{-1}\boldsymbol{D}^2(\boldsymbol{D}^2 + 2\lambda\boldsymbol{I})^{-1}\boldsymbol{V}^T\right) = \\ &= \sigma^2\sum_{j=1}^{p}\frac{d_j^2}{(d_j^2 + 2\lambda)^2}. \end{aligned} \tag{3.47}$$

Cuanto mayor sea el valor de λ, menor será la Variancia Total de $\hat{\boldsymbol{\beta}}^{Ridge}$.

Cuando $\lambda \geq 0$, se verifica que $\frac{d_j^2}{(d_j^2+2\lambda)^2} \leq \frac{1}{d_j^2}$, por lo tanto

$$\text{tr}\left(Cov\left(\hat{\boldsymbol{\beta}}^{Ridge}|\boldsymbol{X}\right)\right) \leq \text{tr}\left(Cov\left(\hat{\boldsymbol{\beta}}^{MC}|\boldsymbol{X}\right)\right), \tag{3.48}$$

es decir, la Variancia Total de $\hat{\boldsymbol{\beta}}^{Ridge}$ siempre es menor o igual que la de $\hat{\boldsymbol{\beta}}^{MC}$.

Los valores ajustados usando regresión *Ridge* resultan

$$\hat{\boldsymbol{y}}^{Ridge} = \boldsymbol{X}\hat{\boldsymbol{\beta}}^{Ridge} = \boldsymbol{H}_\lambda \boldsymbol{y}, \tag{3.49}$$

donde $\boldsymbol{H}_\lambda = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X} + 2\lambda\boldsymbol{I})^{-1}\boldsymbol{X}^T$ es una matriz de proyección.

Usando la DVS de la matriz $\boldsymbol{X}$ (ec. 3.7), la matriz $\boldsymbol{H}_\lambda$ resulta

$$\begin{aligned}\boldsymbol{H}_\lambda &= \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^T(\boldsymbol{V}\boldsymbol{D}^2\boldsymbol{V}^T + 2\lambda\boldsymbol{I})^{-1}\boldsymbol{V}\boldsymbol{D}\boldsymbol{U}^T = \\ &= \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^T\boldsymbol{V}(\boldsymbol{D}^2 + 2\lambda\boldsymbol{I})^{-1}\boldsymbol{V}^T\boldsymbol{V}\boldsymbol{D}\boldsymbol{U}^T = \boldsymbol{U}\boldsymbol{D}(\boldsymbol{D}^2 + 2\lambda\boldsymbol{I})^{-1}\boldsymbol{D}\boldsymbol{U}^T.\end{aligned} \tag{3.50}$$

Dado que la matriz $\boldsymbol{H}_\lambda$ es diagonalizable con respecto a $\boldsymbol{U}$, con autovalores dados por $\boldsymbol{D}(\boldsymbol{D}^2 + 2\lambda\boldsymbol{I})^{-1}\boldsymbol{D}$, y que la traza de una matriz es la suma de sus autovalores, resulta

$$\text{tr}(\boldsymbol{H}_\lambda) = \sum_{j=1}^{p} \frac{d_j^2}{d_j^2 + 2\lambda}. \tag{3.51}$$

Esta función monótona decreciente en λ son los grados de libertad efectivos del ajuste de una regresión *Ridge*. Usualmente, en el ajuste de una regresión lineal con p variables explicativas, el número de grados de libertad del ajuste es igual a p, el número de parámetros libres. En el ajuste de una regresión *Ridge*, a pesar de que generalmente los p coeficientes ajustados serán distintos de cero, la magnitud de estos coeficientes está sujeta a una res-

tricción controlada por el valor de λ. Este es el motivo por el cual se habla de grados de libertad efectivos. Cuando no se aplica regularización ($\lambda = 0$), se tiene que $\text{tr}(\boldsymbol{H}_\lambda) = p$. A medida que λ aumenta, los grados de libertad efectivos se acercan a cero. Por lo tanto, la regularización conduce a una reducción de los mismos.

3.4.2. Regresión LASSO

El estimador LASSO se define como

$$\hat{\boldsymbol{\beta}}^{LASSO} = \underset{\boldsymbol{\beta} \in \mathbb{R}^p}{\text{argmin}} \left\{ \frac{1}{2} \sum_{i=1}^{n} (y_i - \boldsymbol{x}_i^T \boldsymbol{\beta})^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}. \tag{3.52}$$

Al igual que en regresión *Ridge*, LASSO está formulado con respecto a la matriz de predictores estandarizados.

La función a minimizar en la ecuación (3.52) escrita en forma matricial resulta

$$\frac{1}{2} ||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}||_2^2 + \lambda ||\boldsymbol{\beta}||_1. \tag{3.53}$$

A diferencia de la regresión *Ridge*, por lo general LASSO no admite una solución en forma cerrada. El uso de la penalización $||\boldsymbol{\beta}||_1$ hace que la solución LASSO sea no lineal en $\boldsymbol{y}$. La minimización de (3.53) constituye un problema de programación cuadrático, cuya solución puede ser aproximada eficientemente (Hastie et al., 2009).

Cuando la matriz $\boldsymbol{X}$ es ortogonal, es posible obtener la expresión de la solución LASSO en forma cerrada. Re-expresando $\frac{1}{2}||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}||_2^2$ en (3.53) se obtiene $\frac{1}{2}\boldsymbol{y}^T\boldsymbol{y} - \boldsymbol{y}^T\boldsymbol{X}\boldsymbol{\beta} + \frac{1}{2}\boldsymbol{\beta}^T\boldsymbol{\beta}$. Dado que el término $\frac{1}{2}\boldsymbol{y}^T\boldsymbol{y}$ no depende de los coeficientes de interés $\boldsymbol{\beta}$, puede omitirse para la búsqueda del mínimo respecto de $\boldsymbol{\beta}$ dando lugar al siguiente problema de optimización

equivalente a (3.29):

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^p} -\boldsymbol{y}^T\boldsymbol{X}\boldsymbol{\beta} + \frac{1}{2}||\boldsymbol{\beta}||_2^2 + \lambda||\boldsymbol{\beta}||_1. \tag{3.54}$$

Teniendo en cuenta que $\left(\boldsymbol{y}^T\boldsymbol{X}\right)^T = \hat{\boldsymbol{\beta}}^{MC}$, el problema (3.54) puede ser escrito como

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \sum_{j=1}^{p} -\hat{\beta}_j^{MC}\beta_j + \frac{1}{2}\beta_j^2 + \lambda|\beta_j|. \tag{3.55}$$

De este modo, la función que se desea minimizar en (3.54) es ahora la suma de p funciones, cada una correspondiente a un coeficiente β_j, las cuales pueden ser resueltas individualmente. El mínimo en (3.54) será igual a la suma de los mínimos obtenidos en cada uno de los p términos de la sumatoria en (3.55).

Para cada $j \in \{1, ..., p\}$ se busca el valor β_j que minimiza la función

$$\mathcal{L}_j = -\hat{\beta}_j^{MC}\beta_j + \frac{1}{2}\beta_j^2 + \lambda|\beta_j|. \tag{3.56}$$

Si $\hat{\beta}_j^{MC} > 0$, el valor β_j que minimiza (3.56) será necesariamente $\beta_j \geq 0$, ya que con el opuesto de todo valor $\beta_j < 0$ se obtiene un valor menor en la función a minimizar (3.56). Aplicando el mismo razonamiento, si $\hat{\beta}_j^{MC} < 0$, el valor β_j que minimiza (3.56) será necesariamente $\beta_j \leq 0$. A continuación se estudia cada uno de estos escenarios.

Escenario 1: $\hat{\beta}_j^{MC} > 0$.

Como $\hat{\beta}_j^{MC} > 0$, necesariamente $\beta_j \geq 0$, y (3.56) resulta

$$\mathcal{L}_j = -\hat{\beta}_j^{MC}\beta_j + \frac{1}{2}\beta_j^2 + \lambda\beta_j. \tag{3.57}$$

Derivando (3.57) con respecto a β_j e igualando a cero se obtiene que $\beta_j = \hat{\beta}_j^{MC} - \lambda$. Esto es válido únicamente si $\hat{\beta}_j^{MC} - \lambda$ es no negativo, por lo tanto, cuando $\hat{\beta}_j^{MC} > 0$ la solución

LASSO resulta

$$\hat{\beta}_j^{LASSO} = (\hat{\beta}_j^{MC} - \lambda)^+ = sign(\hat{\beta}_j^{MC})(|\hat{\beta}_j^{MC}| - \lambda)^+, \tag{3.58}$$

donde $sign(\hat{\beta}_j^{MC})$ es la función signo aplicada a $\hat{\beta}_j^{MC}$ y $(|\hat{\beta}_j^{MC}| - \lambda)^+$ es la parte positiva de $(|\hat{\beta}_j^{MC}| - \lambda)$.

Escenario 2: $\hat{\beta}_j^{MC} < 0$.

Como $\hat{\beta}_j^{MC} < 0$, necesariamente $\beta_j \leq 0$, y por lo tanto (3.56) se expresa como

$$\mathcal{L}_j = -\hat{\beta}_j^{MC}\beta_j + \frac{1}{2}\beta_j^2 - \lambda\beta_j. \tag{3.59}$$

Derivando (3.59) con respecto a β_j e igualando a cero se obtiene que $\beta_j = \hat{\beta}_j^{MC} + \lambda$. Recordando que esto es válido únicamente si $\hat{\beta}_j^{MC} + \lambda$ es no positivo, la solución LASSO cuando $\hat{\beta}_j^{MC} < 0$ resulta

$$\hat{\beta}_j^{LASSO} = sign(\hat{\beta}_j^{MC})(|\hat{\beta}_j^{MC}| - \lambda)^+. \tag{3.60}$$

El resultado (3.58) obtenido en el **Escenario 1** coincide con el resultado (3.60) obtenido en el **Escenario 2**, por lo tanto, la expresión del estimador LASSO cuando la matriz $\boldsymbol{X}$ es ortogonal resulta

$$\hat{\boldsymbol{\beta}}^{LASSO} = sign(\hat{\boldsymbol{\beta}}^{MC})(|\hat{\boldsymbol{\beta}}^{MC}| - \boldsymbol{\lambda})^+, \tag{3.61}$$

donde $\boldsymbol{\lambda}$ es un vector de coeficientes λ de dimensión $p \times 1$.

Otro modo de expresar el estimador LASSO cuando la matriz $\boldsymbol{X}$ es ortogonal es utilizando una función denominada *soft-thresholding*, la cual se simboliza $S_t(\cdot)$. Al subíndice t se lo llama nivel. Esta función es igual a cero si el valor absoluto de su argumento es menor que el nivel, y si el valor absoluto de su argumento es mayor o igual que el nivel, la función es igual al argumento trasladado hacia cero en una magnitud igual al nivel. Es decir, si su argumento es positivo se resta el nivel, y si es negativo, se suma. De este modo, el estimador LASSO

(3.61) puede escribirse como

$$\hat{\boldsymbol{\beta}}^{LASSO} = S_\lambda \left(\hat{\boldsymbol{\beta}}^{MC} \right). \tag{3.62}$$

Resumiendo:

$$\hat{\beta}_j^{LASSO}(\lambda) = \begin{cases} \hat{\beta}_j^{MC} - \lambda & \text{si } \hat{\beta}_j^{MC} > \lambda \\ 0 & \text{si } |\hat{\beta}_j^{MC}| \leq \lambda \\ \hat{\beta}_j^{MC} + \lambda & \text{si } \hat{\beta}_j^{MC} < -\lambda \end{cases} \tag{3.63}$$

Propiedades de la Regresión LASSO

El problema LASSO en (3.29) no necesariamente tiene una única solución $\hat{\boldsymbol{\beta}}^{LASSO}$, debido a que la función a minimizar no es estrictamente convexa cuando $\boldsymbol{X}^T\boldsymbol{X}$ es singular, hecho que ocurre cuando $p > n$. De todos modos, el conjunto de valores ajustados $\boldsymbol{X}\hat{\boldsymbol{\beta}}^{LASSO}$ sí es único, dado que la SCR es estrictamente convexa en $\boldsymbol{X}\boldsymbol{\beta}$ (Tibshirani et al., 2013).

Las condiciones de Karush-Kuhn-Tucker (Wu, 2007), también conocidas como las condiciones KKT, son un conjunto de condiciones de optimización que caracterizan la solución de cualquier problema convexo. Para el problema LASSO en (3.29), estas condiciones determinan que cualquier solución $\hat{\boldsymbol{\beta}}^{LASSO}$ debe satisfacer

$$\boldsymbol{X}^T(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}^{LASSO}) = \lambda \boldsymbol{s}, \tag{3.64}$$

donde $\boldsymbol{s} \in \partial||\hat{\boldsymbol{\beta}}^{LASSO}||_1$ es un subgradiente de la norma ℓ_1 evaluada en $\hat{\boldsymbol{\beta}}^{LASSO}$, es decir

$$s_j \in \begin{cases} \{+1\} & \text{si } \hat{\beta}_j^{LASSO} > 0 \\ \{-1\} & \text{si } \hat{\beta}_j^{LASSO} < 0 \\ [-1, 1] & \text{si } \hat{\beta}_j^{LASSO} = 0. \end{cases} \quad j = 1, ..., p \tag{3.65}$$

De este modo, $\hat{\boldsymbol{\beta}}^{LASSO}$ es solución de (3.29) si y solo si ese estimador satisface (3.64) y (3.65) para algún $\boldsymbol{s}$.

En (3.64) se observa que, cuando $\lambda > 0$, el subgradiente óptimo $\boldsymbol{s}$ es una función de los valores ajustados $\boldsymbol{X}\hat{\boldsymbol{\beta}}^{LASSO}$, los cuales son determinados de forma única, por lo que $\boldsymbol{s}$ es en sí mismo único. Esto es cierto incluso cuando la solución $\hat{\boldsymbol{\beta}}^{LASSO}$ no está determinada unívocamente.

Por (3.65), la unicidad de $\boldsymbol{s}$ implica que cualquier par de soluciones LASSO debe tener los mismos signos comparando componente a componente. Esto quiere decir que no pueden encontrarse dos soluciones $\hat{\boldsymbol{\beta}}^{LASSO}$ y $\tilde{\boldsymbol{\beta}}^{LASSO}$ con $\hat{\beta}_j^{LASSO} > 0$ y $\tilde{\beta}_j^{LASSO} < 0$ para algún j.

De este modo, las soluciones LASSO no presentan problemas al momento de interpretar los signos de sus componentes.

La solución LASSO será única cuando las columnas de la matriz de predictores $\boldsymbol{X}$ estén en *posición general* (Tibshirani et al., 2013). Se dice que $\boldsymbol{X}_1$, ..., $\boldsymbol{X}_p \in \mathbb{R}^n$ están en posición general si para cualquier $w < \text{mín}\{n, p\}$, índices $i_1, ..., i_{w+1} \in \{1, ..., p\}$ y signos $c_1, ..., c_{w+1} \in \{-1, +1\}$, el espacio generado por $c_1\boldsymbol{X}_{i_1}, ..., c_{w+1}\boldsymbol{X}_{i_{w+1}}$ no contiene elemento alguno de $\{\pm\boldsymbol{X}_i : i \neq i_1, ..., i_{w+1}\}$. Esto es equivalente a decir que ningún subespacio w-dimensional $\Omega \subseteq \mathbb{R}^n$ contiene más de $w+1$ puntos de $\{\pm\boldsymbol{X}_1, ..., \pm\boldsymbol{X}_p\}$, para $w < \text{mín}\{n, p\}$ y excluyendo los pares $+\boldsymbol{X}_i$ y $-\boldsymbol{X}_i$. Básicamente, esta propiedad implica que las columnas de $\boldsymbol{X}$ no presentan dependencias lineales de bajo orden.

Sin importar los tamaños relativos de n y p, las columnas de $\boldsymbol{X}$ estarán en posición general con probabilidad uno siempre que las columnas de la matriz de predictores $\boldsymbol{X}$ sigan una distribución conjunta continua (Tibshirani et al., 2013). Por este motivo, los estimadores LASSO generalmente no presentan el problema de falta de unicidad en grandes dimensiones como sucede con los estimadores mínimo-cuadráticos.

Saturación

Sea A el soporte, también denominado conjunto activo de la solución LASSO. Este conjunto está formado por todos los índices de las variables cuyos coeficientes estimados son distintos de cero. En símbolos: $A = sop\left(\hat{\boldsymbol{\beta}}^{LASSO}\right) = \{j : \hat{\beta}_j^{LASSO} \neq 0\}$.

Cuando $\boldsymbol{X}$ está en posición general, para todo $\lambda > 0$, la submatriz $\boldsymbol{X}_A$, formada por las columnas de la matriz $\boldsymbol{X}$ correspondientes a las variables que forman parte del conjunto activo, siempre tiene rango completo en las columnas. Esto implica que $\#A \leq \text{mín}\{n, p\}$, es decir, la solución LASSO nunca tendrá más de mín$\{n, p\}$ componentes distintas de cero. Esta propiedad recibe el nombre de **saturación**, y no siempre es deseable en la práctica ya que, por ejemplo, si se tuviesen 100.000 variables explicativas continuas y tan solo 100 individuos, usando regresión LASSO nunca se podría construir un modelo lineal con más de 100 predictores.

3.5. Dispersión

En el contexto de grandes dimensiones, el número de parámetros $\boldsymbol{\beta}$ a estimar es muy elevado, y es de esperar que una gran parte de las variables explicativas no influyan en el valor de la respuesta. Generalmente, a menos que se imponga alguna condición adicional en el método de estimación, los estimadores $\hat{\beta}_j$ serán distintos de cero para todo $j \in \{1, ..., p\}$.

Se dice que una solución $\hat{\boldsymbol{\beta}}$ es dispersa cuando $\hat{\beta}_j = 0$ para un gran número de las componentes $j \in \{1, ..., p\}$. La dispersión en las soluciones es una propiedad deseable en un estimador ya que se corresponde con la realización de una selección de variables en el modelo lineal construído, posibilitando la obtención de modelos interpretables.

La solución de la selección del mejor subconjunto $\left(\hat{\boldsymbol{\beta}}^{MS}\right)$ es dispersa. Esto se debe al uso de la norma ℓ_0 para definir al estimador, ya sea en forma restringida o penalizada, la cual se

define justamente como el número de elementos distintos de cero en un vector. En (3.25), el parámetro $r > 0$ es el que controla el nivel de dispersión de la solución. Cuanto más pequeño sea el valor de r, más dispersa es la solución $\hat{\boldsymbol{\beta}}^{MS}$, es decir, el número de $\hat{\beta}_j^{MS} = 0$ es mayor. En (3.28), la solución $\hat{\boldsymbol{\beta}}^{MS}$ es más dispersa cuanto mayor sea el valor de $\lambda \geq 0$.

La solución $\hat{\boldsymbol{\beta}}^{LASSO}$ también verifica esta propiedad. Esto se debe al uso de la norma ℓ_1 en su definición, ya sea en forma restringida o penalizada. En (3.26), cuanto menor sea el valor de $t \geq 0$, más dispersa será la solución. En (3.29), el parámetro $\lambda \geq 0$ es el que controla el nivel de dispersión. La solución $\hat{\boldsymbol{\beta}}^{LASSO}$ es más dispersa cuanto mayor sea el valor de $\lambda \geq 0$.

La solución $\hat{\boldsymbol{\beta}}^{Ridge}$ no es dispersa. Generalmente, $\hat{\boldsymbol{\beta}}^{Ridge}$ tiene todas sus componentes distintas de cero, sin importar el valor de $t \geq 0$ en (3.27) o $\lambda \geq 0$ en (3.30).

Para justificar que el uso de la norma ℓ_1, a diferencia de la norma ℓ_2, induce dispersión, se analizan gráficamente los problemas LASSO y *Ridge*, expresados en su forma restringida, cuando el número de parámetros a estimar p es igual a 2 ($\boldsymbol{\beta} = (\beta_1, \beta_2)$). En la Figura 3.1, el diamante $|\beta_1| + |\beta_2| \leq t$ (ubicado en el sector izquierdo de la figura) y la circunferencia $\beta_1^2 + \beta_2^2 \leq t$ (ubicada en el sector derecho) son las restricciones LASSO y *Ridge* respectivamente. En $\mathbb{R}^2$, la SCR tiene curvas de nivel elípticas (elipses rojas de la Figura 3.1), con centro en el estimador mínimo-cuadrático. Las soluciones LASSO y *Ridge* se encuentran en el punto donde el contorno que corresponde a un nivel particular de la SCR intercepta esas restricciones. A diferencia de la circunferencia, el diamante tiene esquinas, y si una solución ocurre en una esquina, entonces tiene un coeficiente $\hat{\beta}_j$ igual a cero. Cuando la restricción está dada por una circunferencia, es más difícil que alguno de los coeficientes estimados sea exactamente cero. Cuando $p > 2$, el diamante se convierte en un romboide, el cual tiene muchas esquinas, bordes planos y caras, aumentando las oportunidades de que los parámetros estimados sean cero y así el uso de la norma ℓ_1 induce dispersión en las soluciones LASSO.

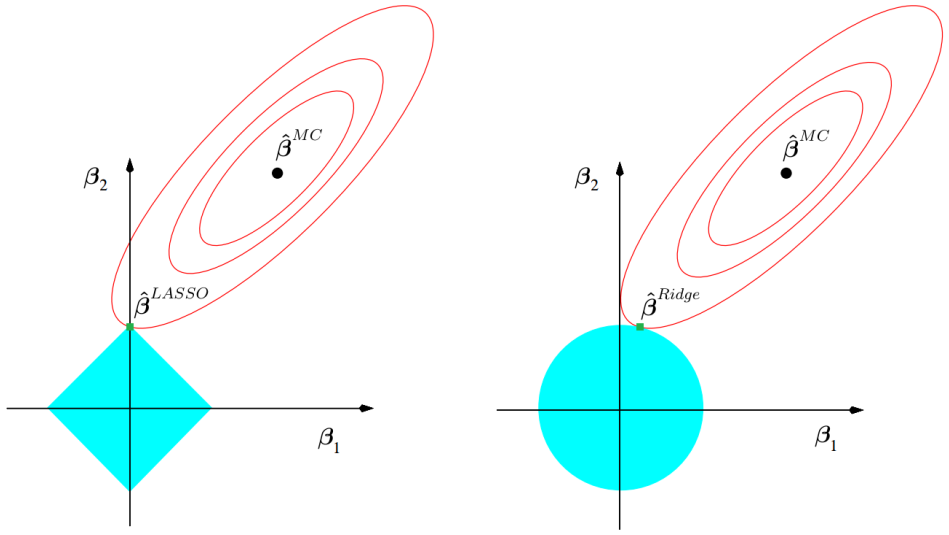


Figura 3.1: Representación geométrica de las condiciones para encontrar los estimadores LASSO (izquierda) y *Ridge* (derecha) cuando $p = 2$. Las áreas sólidas son las regiones de restricción $|\beta_1| + |\beta_2| \leq t$ (LASSO) y $\beta_1^2 + \beta_2^2 \leq t$ (*Ridge*), mientras que las elipses rojas son las curvas de nivel de la SCR. Figura adaptada de *The Elements of Statistical Learning 2nd Edition* (p. 71), por Hastie, T., Tibshirani, R. y Friedman, J., 2009, New York, NY: Springer.

3.6. Optimización Convexa

Los estimadores en las regresiones *Ridge* y LASSO plantean problemas de optimización convexos (ver Anexo I), mientras que la selección del mejor subconjunto es un problema de optimización no convexo de los más complejos.

La propiedad de convexidad es muy útil ya que asegura que el movimiento continuo que produce iterativamente un algoritmo en direcciones que disminuyen la función objetivo alcanzará eventualmente el mínimo global.

Asumiendo $\lambda > 0$, el problema de optimización *Ridge* en (3.30) es estrictamente convexo debido a la presencia de la penalidad $||\boldsymbol{\beta}||_2^2$. Esto es cierto para cualquier matriz de predictores $\boldsymbol{X}$, por lo que la solución *Ridge* siempre está bien definida y, de hecho, está dada en forma cerrada por $\hat{\boldsymbol{\beta}}^{Ridge} = (\boldsymbol{X}^T\boldsymbol{X} + 2\lambda\boldsymbol{I})^{-1}\boldsymbol{X}^T\boldsymbol{y}$.

Por el contrario, debido a la presencia de la penalidad $||\boldsymbol{\beta}||_1$, el problema LASSO en

(3.29) no siempre es estrictamente convexo y por lo tanto, no necesariamente tiene solución única. Es posible definir un problema modificado que siempre es estrictamente convexo, a través de la *red elástica* (Zou and Hastie, 2005):

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \frac{1}{2}||\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}||_2^2 + \lambda||\boldsymbol{\beta}||_1 + \delta||\boldsymbol{\beta}||_2^2, \tag{3.66}$$

donde ahora ambos λ, $\delta > 0$ son parámetros de suavizado. Además de garantizar unicidad para cualquier matriz $\boldsymbol{X}$, la *red elástica* combina algunas de las propiedades predictivas deseables de la regresión *Ridge* con las propiedades de dispersión de la regresión LASSO.

3.7. Camino de Soluciones LASSO

Usando las condiciones KKT (3.64) y (3.65), es posible calcular la solución LASSO como función de λ, para todos los valores del parámetro de suavizado $\lambda \in [0, \infty)$. El conjunto de estimadores $\hat{\boldsymbol{\beta}}^{LASSO}(\lambda)$ recibe el nombre de camino de regularización o camino de soluciones del problema (3.29). En otras palabras, el camino de regularización $\{\hat{\boldsymbol{\beta}}^{LASSO}(\lambda) : \lambda \in [0, \infty)\}$ describe las soluciones LASSO a medida que varía el parámetro de suavizado λ.

Este camino de soluciones es una función continua lineal por tramos en λ. La continuidad del camino de soluciones permite que sólo sea necesario calcular y almacenar ciertos nodos ($\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_r \geq 0$) y la solución LASSO en los mismos, a partir de los cuales por interpolación lineal pueden calcularse las soluciones para cualquier $\lambda \geq 0$.

Los nodos $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_r$ del camino de soluciones corresponden a valores de λ en los que el conjunto activo $A(\lambda) = sop\left(\hat{\boldsymbol{\beta}}^{LASSO}(\lambda)\right)$ cambia. A medida que λ disminuye desde ∞ a 0, los nodos usualmente corresponderán a puntos en los que una variable ingresa al conjunto activo. Sin embargo, a medida que λ disminuye, un nodo del camino LASSO puede también corresponder a un punto en el cual una variable abandona el conjunto activo, siendo

este el motivo por el cual la regresión LASSO no produce modelos anidados y el número de nodos r puede ser mucho mayor que p. En la Figura 3.2 se presenta un ejemplo del camino LASSO. Cada línea coloreada representa una componente de la solución LASSO $\hat{\beta}_j^{LASSO}(\lambda)$, $j = 1, ..., p$ como función de λ. Las líneas de puntos verticales de color gris marcan los nodos $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_{11}$.

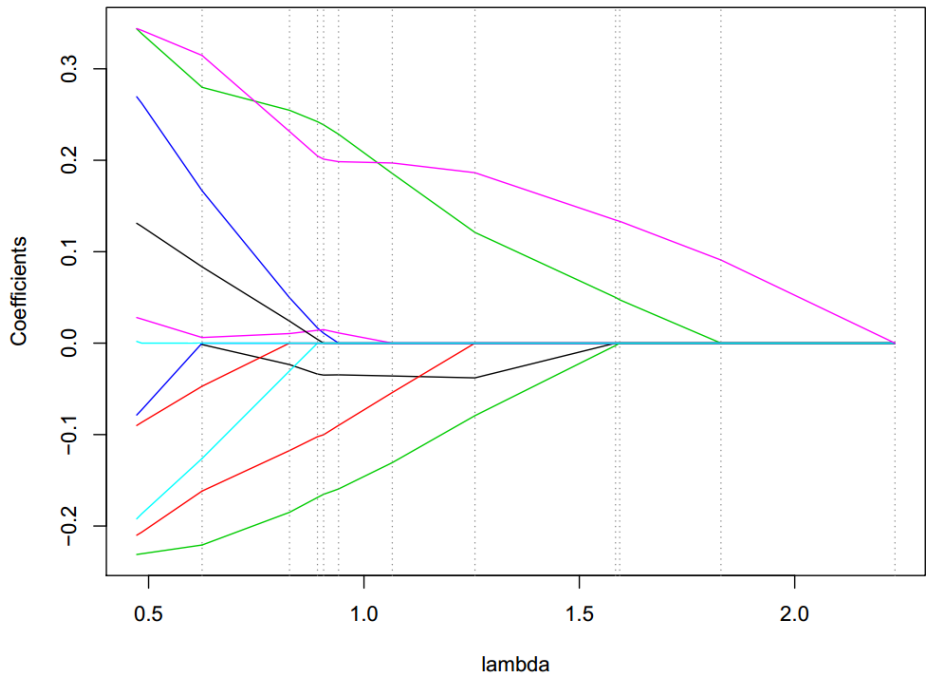


Figura 3.2: Ejemplo de un camino de soluciones LASSO con 11 nodos

El camino de soluciones LASSO fue descripto por Osborne et al. (2000) y Efron et al. (2004) y se construye a partir de la verificación iterativa de las condiciones KKT.

Inicialmente se considera que λ decrece desde ∞ y se define $\hat{\boldsymbol{\beta}}^{LASSO}(\lambda) = \mathbf{0}$. Para que efectivamente esta sea una solución LASSO válida, debe verificar la condición (3.64), es decir, debe existir algún subgradiente $\boldsymbol{s}$ válido para esa solución. Luego de reemplazar $\hat{\boldsymbol{\beta}}^{LASSO}(\lambda) = \mathbf{0}$ en (3.64), resulta $\boldsymbol{X}^T\boldsymbol{y} = \lambda \boldsymbol{s}$, donde $\boldsymbol{s}$ es un subgradiente de la norma ℓ_1 evaluada en $\hat{\boldsymbol{\beta}}^{LASSO}(\lambda) = \mathbf{0}$, es decir, $s_j \in [-1, 1]$ para todo $j = 1, ..., p$. Para valores de λ lo suficientemente grandes, esta implicación es cierta, ya que se puede elegir $\boldsymbol{s} = \boldsymbol{X}^T\boldsymbol{y}/\lambda$. Sin embargo, cuando λ disminuye habiendo pasado el punto en el que $\lambda = |\boldsymbol{X}_j^T\boldsymbol{y}|$ para algún

$j = 1, ..., p$, $\boldsymbol{s}$ deja de ser un subgradiente válido para ese j, ya que $s_j > 1$ para $\hat{\beta}_j^{LASSO} = 0$, lo cual no verifica las condiciones KKT. Resumiendo, $\hat{\boldsymbol{\beta}}^{LASSO}(\lambda) = \mathbf{0}$ es la solución LASSO para todo $\lambda \geq \lambda_1$, donde

$$\lambda_1 = \max_{j=1,...,p} |\boldsymbol{X}_j^T \boldsymbol{y}|. \tag{3.67}$$

Una vez alcanzado el primer nodo λ_1, cuando λ decrece desde ese valor resulta necesario modificar $\hat{\boldsymbol{\beta}}^{LASSO}(\lambda)$ para que las condiciones KKT permanezcan satisfechas. Sea j_1 el índice de la variable que alcanza el máximo en (3.67). Dado que el subgradiente tiene $|s_{j_1}| = 1$ en $\lambda = \lambda_1$, se tiene "permitido" hacer que $\hat{\beta}_{j_1}$ sea distinto de cero. Entonces, a medida que λ decrece desde λ_1, se considera fijar

$$\begin{aligned} &\hat{\beta}_{j_1}^{LASSO}(\lambda) = (\boldsymbol{X}_{j_1}^T \boldsymbol{X}_{j_1})^{-1}(\boldsymbol{X}_{j_1}^T \boldsymbol{y} - \lambda s_{j_1}), \\ &\hat{\beta}_j^{LASSO}(\lambda) = 0, \text{ para todo } j \neq j_1, \end{aligned} \tag{3.68}$$

donde $s_{j_1} = sign(\boldsymbol{X}_{j_1}^T \boldsymbol{y})$. Esta consideración hace que $\hat{\boldsymbol{\beta}}^{LASSO}(\lambda)$ sea una función lineal por tramos continua en λ. El conjunto activo A en este momento resulta $A = \{j_1\}$, mientras que $\boldsymbol{s}_A$, el conjunto que contiene los signos de los coeficientes estimados de las variables en A, también denominado conjunto de signos activos, resulta $\boldsymbol{s}_A = (s_{j_1})$.

Debido a que $\boldsymbol{X}\hat{\boldsymbol{\beta}}^{LASSO} = \boldsymbol{X}_{j_1}\hat{\beta}_{j_1}^{LASSO}$ por ser $\hat{\beta}_j^{LASSO} = 0$ para todo $j \neq j_1$, las condiciones KKT para la variable j_1 son

$$\boldsymbol{X}_{j_1}^T \left(\boldsymbol{y} - \boldsymbol{X}_{j_1}(\boldsymbol{X}_{j_1}^T \boldsymbol{X}_{j_1})^{-1}(\boldsymbol{X}_{j_1}\boldsymbol{y} - \lambda s_{j_1})\right) = \lambda s_{j_1}, \tag{3.69}$$

y para todo $j \neq j_1$

$$|\boldsymbol{X}_j^T \left(\boldsymbol{y} - \boldsymbol{X}_{j_1}(\boldsymbol{X}_{j_1}^T \boldsymbol{X}_{j_1})^{-1}(\boldsymbol{X}_{j_1}\boldsymbol{y} - \lambda s_{j_1})\right)| \leq \lambda. \tag{3.70}$$

En (3.70) se mantiene una desigualdad estricta cuando $\lambda = \lambda_1$ para todo $j \neq j_1$ y, por la continuidad de la solución $\hat{\boldsymbol{\beta}}^{LASSO}(\lambda)$ construida, esto se mantendrá cierto hasta que uno de los tramos lineales $\boldsymbol{X}_j^T(\boldsymbol{y} - \boldsymbol{X}_{j_1}(\boldsymbol{X}_{j_1}^T\boldsymbol{X}_{j_1})^{-1}(\boldsymbol{X}_{j_1}^T\boldsymbol{y} - \lambda s_{j_1}))$, con $j \neq j_1$, resulte igual a $\pm\lambda$, punto en el cual se deberá modificar nuevamente la solución, ya que de otro modo el gradiente implícito $s_j = \dfrac{\boldsymbol{X}_j^T(\boldsymbol{y} - \boldsymbol{X}_{j_1}(\boldsymbol{X}_{j_1}^T\boldsymbol{X}_{j_1})^{-1}(\boldsymbol{X}_{j_1}^T\boldsymbol{y} - \lambda s_{j_1}))}{\lambda}$ dejará de pertenecer al intervalo $[-1, 1]$.

La linealidad del camino de soluciones permite calcular explícitamente el momento en el que uno de los tramos lineales se resulta igual a $\pm\lambda$, el cual recibe el nombre de *hitting time*. A la variable que produce esta igualdad se la denominará variable *hitting*.

La solución LASSO se mantiene igual a (3.68) para todo $\lambda_1 \geq \lambda \geq \lambda_2$, donde

$$\lambda_2 = \max_{j \neq j_1, s_j \in \{-1,1\}}^{+} \frac{\boldsymbol{X}_j^T(\boldsymbol{I} - \boldsymbol{X}_{j_1}(\boldsymbol{X}_{j_1}^T\boldsymbol{X}_{j_1})^{-1}\boldsymbol{X}_{j_1})\boldsymbol{y}}{s_j - \boldsymbol{X}_j^T\boldsymbol{X}_{j_1}(\boldsymbol{X}_{j_1}^T\boldsymbol{X}_{j_1})^{-1}s_{j_1}}, \tag{3.71}$$

siendo $\max^+$ el máximo sobre todos los argumentos que son menores que λ_1.

Ahora bien, sean $A = \{j_1, j_2\}$ y $\boldsymbol{s}_A = (s_{j_1}, s_{j_2})$, donde j_2 y s_2 son la variable y el signo que alcanzan el máximo en (3.71). A medida que λ decrece desde λ_2, se considera fijar

$$\begin{aligned} \hat{\boldsymbol{\beta}}_A^{LASSO}(\lambda) &= (\boldsymbol{X}_A^T\boldsymbol{X}_A)^{-1}(\boldsymbol{X}_A^T\boldsymbol{y} - \lambda\boldsymbol{s}_A), \\ \hat{\boldsymbol{\beta}}_{-A}^{LASSO}(\lambda) &= \boldsymbol{0}, \end{aligned} \tag{3.72}$$

donde $\hat{\boldsymbol{\beta}}_A^{LASSO}$ se refiere a los coeficientes estimados de las variables del conjunto activo, mientras que $\hat{\boldsymbol{\beta}}_{-A}^{LASSO}$ se refiere a los de las variables que no forman parte de dicho conjunto.

Nuevamente, las condiciones KKT se verificarán para un conjunto de valores de λ, pero cuando uno de los $\boldsymbol{X}_j^T(\boldsymbol{y} - \boldsymbol{X}_A(\boldsymbol{X}_A^T\boldsymbol{X}_A)^{-1}(\boldsymbol{X}_A^T\boldsymbol{y} - \lambda\boldsymbol{s}_A))$, $j \notin A$, resulte igual a $\pm\lambda$, se deberá volver a modificar el estimador $\hat{\boldsymbol{\beta}}^{LASSO}$ para que se satisfagan dichas condiciones.

Al disminuir λ cuando el conjunto activo está formado por dos o más variables, además de

tener en cuenta el siguiente *hitting time*, es necesario verificar que ninguna de las componentes activas de la solución calculada en (3.72) cruce el valor cero, porque pasado ese punto, $\boldsymbol{s}_A$ dejará de ser un subgradiente válido para esas componentes activas. A modo de ejemplo, si $\hat{\beta}_j^{LASSO}$ pasa de positivo a negativo, $s_j = 1$ dejará de ser un subgradiente válido para ese j.

El momento en el que una componente $\hat{\beta}_j^{LASSO}$ cambia de signo recibe el nombre de *crossing time*, y también puede ser calculado explícitamente debido a la linealidad del camino de soluciones. A la variable cuya componente cambia de signo se la denominará variable *crossing*.

Por lo tanto, se sostiene que (3.72) es la solución LASSO para todo $\lambda_2 \geq \lambda \geq \lambda_3$, donde λ_3 es el máximo entre el siguiente *hitting time* y *crossing time*. Si λ_3 corresponde a un *hitting time*, la variable *hitting* se agrega al conjunto activo. Por el contrario, si corresponde a un *crossing time*, se remueve la variable *crossing* de dicho conjunto. Cualquiera sea el caso, es necesario volver a calcular la solución LASSO usando (3.72) de modo que se verifiquen las condiciones KKT. Luego, se calculan los siguientes *hitting* y *crossing times*, se verifica si es necesario agregar o eliminar una variable al conjunto activo y vuelve a calcularse la solución LASSO. Este procedimiento se repite mientras $\lambda > 0$. El algoritmo del camino LASSO se resume a continuación:

Algoritmo 1. Algoritmo del Camino de Soluciones LASSO

Dados $\boldsymbol{X}$ e $\boldsymbol{y}$

- Comenzar con el contador de iteración $k = 0$, parámetro de regularización $\lambda_0 = \infty$, conjunto activo $A = \emptyset$ y signos activos $\boldsymbol{s}_A = \emptyset$.
- Mientras $\lambda_k > 0$:
    1. Calcular la solución LASSO a medida que λ decrece desde λ_k a través de

$$\hat{\boldsymbol{\beta}}_A(\lambda) = (\boldsymbol{X}_A^T \boldsymbol{X}_A)^{-1}(\boldsymbol{X}_A^T \boldsymbol{y} - \lambda \boldsymbol{s}_A),$$
$$\hat{\boldsymbol{\beta}}_{-A}(\lambda) = \boldsymbol{0},$$

2. Calcular el siguiente *hitting time*

$$\lambda_{k+1}^{hit} = \max_{j \neq A, s_j \in \{-1,1\}}^{+} \frac{\boldsymbol{X}_j^T(\boldsymbol{I} - \boldsymbol{X}_A(\boldsymbol{X}_A^T\boldsymbol{X}_A)^{-1}\boldsymbol{X}_A)\boldsymbol{y}}{s_j - \boldsymbol{X}_j^T\boldsymbol{X}_A(\boldsymbol{X}_A^T\boldsymbol{X}_A)^{-1}\boldsymbol{s}_A},$$

donde $\max^+$ denota el máximo de los argumentos menores que λ_k.

3. Calcular el siguiente *crossing time*

$$\lambda_{k+1}^{cross} = \max_{j \in A}^{+} \frac{[(\boldsymbol{X}_A^T\boldsymbol{X}_A)^{-1}\boldsymbol{X}_A)\boldsymbol{y}]_j}{[(\boldsymbol{X}_A^T\boldsymbol{X}_A)^{-1}\boldsymbol{s}_A]_j},$$

4. Decrecer λ hasta λ_{k+1}, definido por

$$\lambda_{k+1} = \max\{\lambda_{k+1}^{hit}, \lambda_{k+1}^{cross}\}$$

5. Si $\lambda_{k+1}^{hit} > \lambda_{k+1}^{cross}$ agregar la variable *hitting* al conjunto A y su signo a $\boldsymbol{s}_A$; caso contrario, remover la variable *crossing* de A y su signo de $\boldsymbol{s}_A$. Actualizar $k = k+1$.

El algoritmo del camino de soluciones LASSO presentado en el Algoritmo 1 también es llamado Regresión del Menor Ángulo en modo LASSO debido a su estrecha relación con otro algoritmo conocido como Regresión del Menor Ángulo o LARS (*Least Angle Regression*), el cual fue presentado por Efron et al. (2004) y también es utilizado para el ajuste de modelos lineales en grandes dimensiones. La única diferencia radica en que LARS no evita que las componentes de las soluciones crucen el cero.

Otro algoritmo que puede ser utilizado como alternativa para el cálculo del camino de soluciones LASSO es el denominado *Coordinate Descent* (Friedman et al., 2007) (Friedman et al., 2010). Este algoritmo mantiene fijo el parámetro de penalidad λ en (3.52) y optimiza sucesivamente sobre cada parámetro β_j $j = 1, ..., p$ hasta lograr la convergencia del proceso. Esto se repite para distintos valores de λ, obteniendo el camino de soluciones LASSO. *Coordinate Descent* es más rápido que LARS, especialmente en problemas de gran magnitud. Además, puede proporcionar soluciones para un conjunto determinado de valores de λ, a diferencia de LARS que necesariamente calcula todo el camino de soluciones.

Este algoritmo se encuentra disponible en forma gratuita en el paquete **glmnet** de

MATLAB o R (Friedman et al., 2009) y es el que ha sido utilizado en esta tesina. En el Anexo II se describe el uso del paquete en R, con algunas instrucciones y particularidades del mismo.

En el Anexo III se presentan programas disponibles para la obtención de los estimadores de los métodos estudiados.

Capítulo 4

Resultados

En este capítulo se presentan los resultados de un estudio por simulación para comparar las propiedades de los estimadores *Ridge* y LASSO en diferentes escenarios. Los estimadores mínimo-cuadráticos son incluidos en la comparación sólo en los casos donde existen estimaciones únicas.

Dado que los métodos *Ridge* y LASSO mejoran las estimaciones obtenidas en presencia de muchos parámetros no significativos en un modelo de regresión lineal, los escenarios previstos consideran la situación extrema de existencia de parámetros nulos, para identificar si el método de estimación los reconoce como tal.

La eficiencia de los modelos se evalúa a través del Error Cuadrático Medio (ECM) y las propiedades de los estimadores a través de su distribución empírica.

Para las simulaciones se utiliza el paquete **simulator** del software R. Este es un paquete que agiliza el proceso de realizar simulaciones al crear una infraestructura común que puede usarse y reutilizarse fácilmente (Bien, 2016). El **simulator** divide la simulación en cuatro componentes:

1. **Modelo**: el modelo estadístico que determina cómo se generan los datos.

2. **Métodos**: los procedimientos estadísticos que se desean comparar. Dados los datos, cada método produce una salida en forma de estimador, predicción o decisión.

3. **Métricas**: medidas que evalúan el desempeño de los métodos utilizando las salidas obtenidas a partir de los datos.

4. **Gráficos**: representación gráfica de las métricas evaluadas para los distintos métodos bajo los distintos escenarios simulados.

El paquete utiliza el código de las primeras tres componentes para realizar la simulación, mientras que la cuarta componente permite presentar los resultados del estudio. En particular, siguiendo el ejemplo "*Betting on sparsity with the LASSO*" (Bien, 2016), en este trabajo se adoptó un modelo lineal esparcido y se calcularon los estimadores LASSO (a través del paquete **glmnet**) y los estimadores *Ridge* (a través de la codificación de (3.37)). Los estimadores MC también se obtuvieron utilizando el paquete **glmnet**.

En la Sección 4.1 se explicita el diseño del estudio por simulación. En la Sección 4.2 se compara la capacidad predictiva de los estimadores. En las Secciones 4.3 y 4.4 se estudian las propiedades distribucionales de los estimadores de parámetros no nulos y nulos respectivamente. Finalmente, en la Sección 4.5 se estudia la capacidad del método LASSO de obtener estimaciones nulas, tanto cuando $\beta = 0$ (situación deseable) como cuando $\beta = 1$ (situación no deseable).

4.1. Diseño del Estudio por Simulación

Se plantearon tres situaciones variando el número de variables explicativas en función del tamaño de muestra, el cual fue fijado en $n = 100$. Las variantes elegidas para el número de predictores fueron $p = n = 100$ (Situación 1), $p = 2n = 200$ (Situación 2) y $p = 4n = 400$ (Situación 3). Para cada combinación de n y p se consideraron modelos lineales múltiples

esparcidos, es decir, modelos con gran cantidad de parámetros iguales a cero, a partir de los cuales se simularon valores de una variable respuesta y. El modelo supuesto es:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}_{npk} + \phi\boldsymbol{\epsilon}, \tag{4.1}$$

siendo,

- $\boldsymbol{y}$ el vector de respuestas de dimensión 100×1
- $\boldsymbol{X}$ una matriz de predictores de dimensión $100 \times p$ cuyos elementos fueron simulados de una distribución $N(0, 1)$
- $\boldsymbol{\beta}_{npk} = \boldsymbol{\beta}_{\gamma}$ un vector de parámetros de dimensión $p \times 1$, a cuyas k primeras componentes se les asignó un valor igual a uno, y a las $p - k$ restantes el valor cero. El valor de k refleja el grado de esparcimiento del modelo
- $\boldsymbol{\epsilon}$ un vector de errores aleatorios de dimensión 100×1 cuyas componentes fueron simuladas de una distribución $N(0, 1)$
- ϕ un parámetro de variabilidad que se supuso $\phi = \sqrt{\frac{\sum_{i=1}^{100} \left(\boldsymbol{X}\boldsymbol{\beta}_{npk}\right)_i^2}{snr \times 100}}$, tomando el *signal-to-noise-ratio* (snr) igual a 2. La elección de este parámetro de variabilidad permite conseguir variantes asociadas al grado de esparcimiento (Bien, 2016)

La matriz de predictores $\boldsymbol{X}$ utilizada en la simulación de valores de la respuesta $\boldsymbol{y}$ se mantuvo invariante para los distintos modelos dentro de cada una de las tres situaciones.

Se consideran variantes para k desde 2 hasta $\frac{p}{2}$, donde a medida que k aumenta, se tiene un mayor número de parámetros distintos de cero, lo que disminuye el grado de esparcimiento del modelo.

En la Situación 1 se construyeron un total de 13 modelos, eligiendo los siguientes valores no equidistantes de k: 2, 4, 6, 8, 10, 15, 20, 25, 30, 35, 40, 45 y 50.

En la Situación 2 se construyeron un total de 18 modelos, tomando como valores de k a las 13 variantes de la Situación 1 y agregando los valores 60, 70, 80, 90 y 100.

En la Situación 3, se construyeron un total de 11 modelos, eligiendo los siguientes valores para k: 2, 5, 10, 20, 30, 40, 50, 75, 100, 150 y 200.

Cada combinación de n, p y k se considera como un escenario donde se compara la eficiencia de los métodos ($\gamma = 1, ..., 42$).

En cada uno de los 42 escenarios definidos se realizaron 1.000 simulaciones de $\boldsymbol{y}$ con los que se obtuvieron los estimadores de $\boldsymbol{\beta}$ correspondientes a las regresiones *Ridge* y LASSO para 50 valores diferentes del parámetro de suavizado λ, reteniendo en cada método aquella solución $\hat{\boldsymbol{\beta}}_\lambda$ que produjera el menor ECM.

En los 13 escenarios de la Situación 1, donde el número de observaciones es igual al número de variables explicativas, también se obtuvo el estimador mínimo-cuadrático. En los otros casos no existe solución única al problema de optimización mínimo-cuadrático, por lo cual no tiene sentido incluirlos en la comparación.

4.1.1. Criterios de Comparación entre Modelos

Para comparar los modelos se tuvieron en cuenta dos criterios: la capacidad predictiva y las propiedades de los estimadores. Para el estudio de las propiedades de los estimadores se construyeron las distribuciones empíricas agrupando aquéllos que estiman parámetros $\beta = 0$ ($\hat{\beta}^{(0)}$) y los que estiman parámetros $\beta = 1$ ($\hat{\beta}^{(1)}$).

Dado que los estimadores LASSO son conocidos por su capacidad de estimar como nulos a una gran cantidad de los parámetros del modelo, lo cual es deseable en el contexto de grandes dimensiones donde el número de variables explicativas puede ser excesivo, se enfoca como medida de eficiencia a las probabilidades que estos estimadores nulos estén reconociendo efectivamente o no a un parámetro sin efecto. En particular, se proponen dos medidas

específicas que reflejan la capacidad del método LASSO de obtener estimaciones nulas, tanto para el caso $\beta = 1$ (situación no deseable) como $\beta = 0$ (situación deseable).

Las medidas consideradas en las comparaciones son:

1. Promedio de Errores Cuadráticos Medios ($\overline{ECM}$)

   Para cada combinación de n, p y k (γ), se calcula para cada método q, $q \in$ {MC, *Ridge*, LASSO} o {*Ridge*, LASSO} según el escenario, el promedio de los ECM entre las 1.000 simulaciones.

$$\overline{ECM}_{q\gamma} = \frac{1}{1000}\sum_{i=1}^{1000} ECM_{q\gamma i} = \frac{1}{1000}\sum_{i=1}^{1000} \frac{||\hat{\boldsymbol{\beta}}_{q\gamma i} - \boldsymbol{\beta}_{\gamma}||_2^2}{p} \tag{4.2}$$

2. Promedio de los estimadores de parámetros $\beta = 1$ y $\beta = 0$

   Para cada método $q \in$ {MC, *Ridge*, LASSO} se define el promedio de los estimadores $\hat{\beta}^{(1)}$ como

$$\overline{\hat{\beta}}^{(1)}_{q\gamma} = \frac{1}{k \times 1000}\sum_{i=1}^{1000}\sum_{l=1}^{k} \hat{\beta}_{q\gamma li} \tag{4.3}$$

   mientras que el promedio de los estimadores $\hat{\beta}^{(0)}$ resulta

$$\overline{\hat{\beta}}^{(0)}_{q\gamma} = \frac{1}{(p-k) \times 1000}\sum_{i=1}^{1000}\sum_{l=k+1}^{p} \hat{\beta}_{q\gamma li} \tag{4.4}$$

3. Variabilidad de las distribuciones muestrales de los estimadores de $\beta = 1$ y $\beta = 0$

   Para cada método $q \in$ {MC, *Ridge*, LASSO} se define el desvío estándar de los estimadores $\hat{\beta}^{(1)}$ como

$$S^{(1)}_{\hat{\beta}_{q\gamma}} = \sqrt{\frac{1}{k \times 1000 - 1}\sum_{i=1}^{1000}\sum_{l=1}^{k}\left(\hat{\beta}_{q\gamma li} - \overline{\hat{\beta}}^{(1)}_{q\gamma}\right)^2}, \tag{4.5}$$

mientras que el desvío estándar de los estimadores de parámetros $\beta = 0$ resulta

$$S^{(0)}_{\hat{\beta}_{q\gamma}} = \sqrt{\frac{1}{(p-k)\times 1000 - 1}\sum_{i=1}^{1000}\sum_{l=k+1}^{p}\left(\hat{\beta}_{q\gamma li} - \overline{\hat{\beta}}^{(0)}_{q\gamma}\right)^2} \tag{4.6}$$

4. Porcentaje de parámetros nulos estimados por cero por el método LASSO

   Esta medida se calcula multiplicando por 100 al cociente entre el número de estimadores de $\beta = 0$ que son 0 y el número de parámetros $\beta = 0$. En símbolos:

$$\left(\frac{1}{(p-k)\times 1000}\sum_{i=1}^{1000}\sum_{l=k+1}^{p} I(\hat{\beta}_{q\gamma li} = 0)\right) \times 100, \tag{4.7}$$

   donde $q = LASSO$. La expresión 4.7 se puede asimilar al concepto de una sensibilidad de los estimadores LASSO para detectar los parámetros no significativos del modelo.

5. Porcentaje de parámetros no nulos estimados por cero por el método LASSO

   Esta medida se calcula multiplicando por 100 al cociente entre el número de estimadores de $\beta = 1$ que son 0 y el número de parámetros $\beta = 1$. En símbolos:

$$\left(\frac{1}{k\times 1000}\sum_{i=1}^{1000}\sum_{l=1}^{k} I(\hat{\beta}_{q\gamma li} = 0)\right) \times 100, \tag{4.8}$$

   donde $q = LASSO$. La expresión 4.8 puede asimilarse a una probabilidad de error de la regresión LASSO en el sentido de un falso positivo.

4.2. Capacidad Predictiva de los Estimadores. Comparación del $\overline{ECM}$.

En esta Sección se evalúa, para cada combinación de n y p, cómo se comportan los $\overline{ECM}$ de cada método a medida que varía el grado de esparcimiento (k).

Situación 1 ($p = n = 100$)

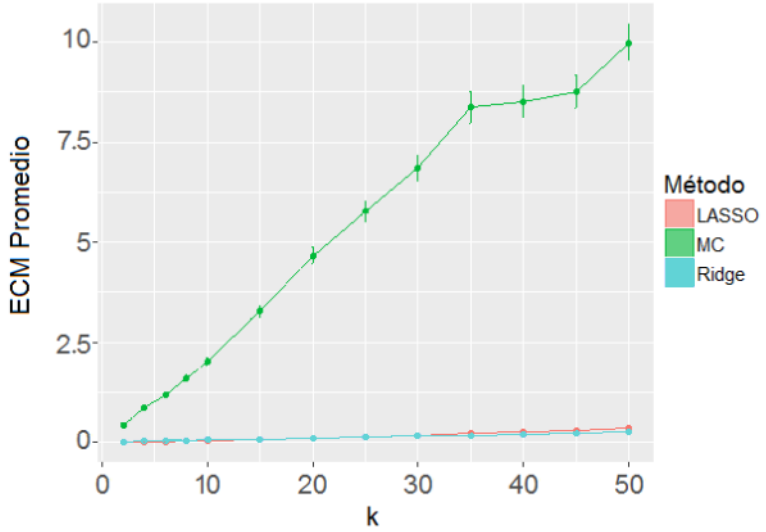


Figura 4.1: $\overline{ECM}$ de los métodos MC, *Ridge* y LASSO para variantes de k. Caso $p = n = 100$

El $\overline{ECM}$ de los estimadores mínimo-cuadráticos aumenta en forma severa a medida que k crece, mientras que el de los métodos de regularización presenta la misma tendencia pero con un crecimiento mucho más lento (Figura 4.1). Esta figura, debido a las dimensiones de las escalas, no muestra diferencias entre los $\overline{ECM}$ de los estimadores *Ridge* y LASSO, sin embargo, no son iguales. Para poder comparar estos dos métodos, se presenta la Figura 4.2 que no considera al método de mínimos cuadrados y se concentra en los valores más pequeños del eje de las ordenadas, evidenciándose las diferencias en el comportamiento de los métodos de regularización.

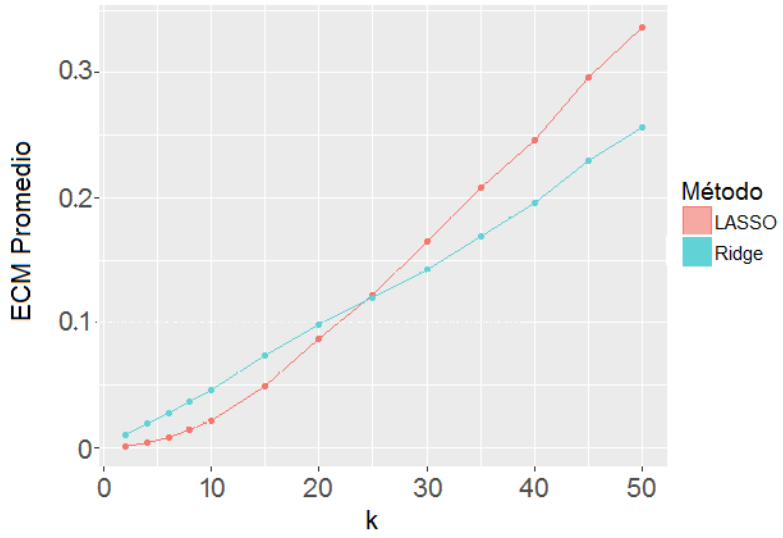


Figura 4.2: $\overline{ECM}$ de los métodos *Ridge* y LASSO para variantes de k. Caso $p = n = 100$

Para valores de k pequeños, menores que 25, el $\overline{ECM}$ de los estimadores LASSO es menor que el de los estimadores *Ridge* (Figura 4.2). Sin embargo, a medida que k se acerca a 25, la diferencia entre los métodos disminuye. A partir de este valor, el $\overline{ECM}$ de los estimadores *Ridge* es menor que el de los LASSO, y la diferencia entre ellos aumenta con k. De todos modos, con valores de k grandes ambos métodos empeoran su desempeño con respecto a los resultados obtenidos para valores de k pequeños.

Situación 2 ($p = 2n = 200$)

En la Figura 4.3 se observa nuevamente que LASSO tiene un mejor desempeño que *Ridge* en cuanto al $\overline{ECM}$ para modelos con mayor grado de esparcimiento, y que esta relación se revierte en problemas más densos. Del mismo modo, con valores de k grandes ambos métodos empeoran su desempeño con respecto a los resultados obtenidos para valores de k pequeños. En este caso, en $k = \frac{p}{2}$ se observan $\overline{ECM}$ mayores que los observados para esa misma proporción de k en la Situación 1 (Figura 4.2).

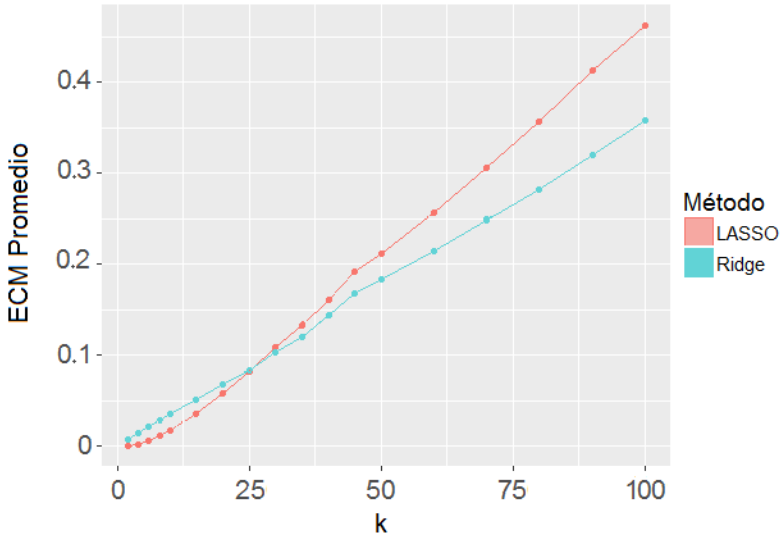


Figura 4.3: $\overline{ECM}$ de los métodos *Ridge* y LASSO para variantes de k. Caso $p = 2n = 200$

Situación 3 ($p = 4n = 400$)

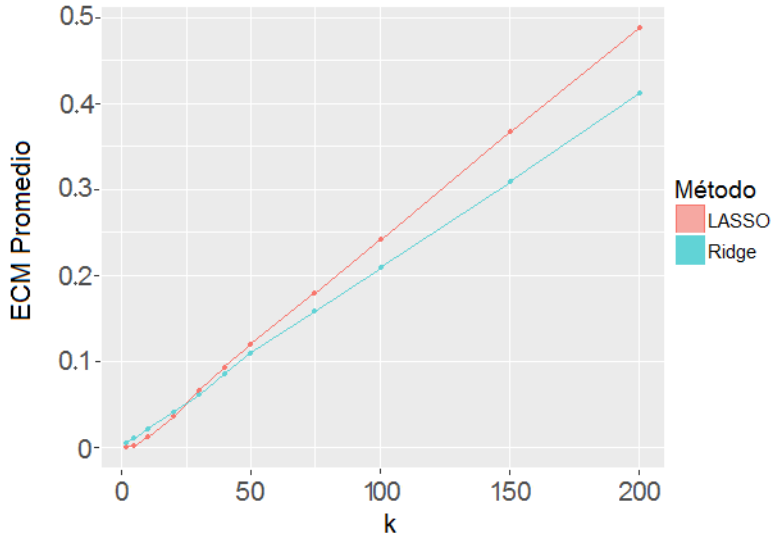


Figura 4.4: $\overline{ECM}$ de los métodos *Ridge* y LASSO para variantes de k. Caso $p = 4n = 400$

En la Situación 3, el comportamiento respecto de la capacidad predictiva de *Ridge* y LASSO (Figura 4.4) repite lo observado en las Situaciones 1 y 2. Los $\overline{ECM}$ aumentan con k y las curvas se cruzan nuevamente aproximadamente en el valor $k = 25$.

4.3. Propiedades Distribucionales de Estimadores $\hat{\beta}^{(1)}$

En esta Sección se comparan los promedios y coeficientes de variación de los estimadores de los parámetros $\beta = 1$ para cada método en cada escenario, separando nuevamente los resultados por situación. El cálculo de los coeficientes de variación es necesario para comparar las variabilidades de las distribuciones frente a promedios diferentes. Además, para completar la descripción del comportamiento aleatorio de los estimadores, se muestran gráficamente sus distribuciones empíricas para cada método. Se han seleccionado sólo algunos escenarios debido a la regularidad de los resultados encontrados. Los casos restantes pueden consultarse en el Anexo IV.

Situación 1 ($p = n = 100$)

Tabla 4.1: Medidas características de las distribuciones de los estimadores $\hat{\beta}^{(1)}$ para cada método y variantes de k. Caso $p = n = 100$.

k	**Método**								
	MC			**Ridge**			**LASSO**		
	Promedio $\overline{\hat{\beta}}^{(1)}_{MC\,\gamma}$	**Desvío Estándar** $S^{(1)}_{\hat{\beta}_{MC\,\gamma}}$	**CV(%)**	**Promedio** $\overline{\hat{\beta}}^{(1)}_{Ridge\,\gamma}$	**Desvío Estándar** $S^{(1)}_{\hat{\beta}_{Ridge\,\gamma}}$	**CV(%)**	**Promedio** $\overline{\hat{\beta}}^{(1)}_{LASSO\,\gamma}$	**Desvío Estándar** $S^{(1)}_{\hat{\beta}_{LASSO\,\gamma}}$	**CV(%)**
2	0,9928	0,6744	67,93	0,5122	0,0739	14,43	0,8197	0,0908	11,08
4	1,0057	0,9940	98,84	0,5160	0,1020	19,77	0,7704	0,1427	18,52
6	1,0089	1,0369	102,78	0,5437	0,1229	22,60	0,7480	0,1711	22,87
8	1,0043	1,1528	114,79	0,5430	0,1525	28,08	0,7209	0,2091	29,01
10	0,9958	1,4126	141,86	0,5346	0,1595	29,84	0,7052	0,2371	33,62
15	0,9828	1,7178	174,79	0,5007	0,1863	37,21	0,6515	0,3269	50,18
20	0,9890	1,9890	201,11	0,4970	0,2184	43,94	0,6077	0,3978	65,46
25	0,9981	2,2679	227,22	0,5241	0,2469	47,11	0,5969	0,4375	73,30
30	0,9942	2,6416	265,70	0,5268	0,2701	51,27	0,5686	0,4775	83,98
35	0,9922	2,8652	288,77	0,5171	0,2966	57,36	0,5381	0,5032	93,51
40	0,9786	2,9323	299,64	0,5151	0,3192	61,97	0,5178	0,5111	98,71
45	0,9792	3,0316	309,60	0,5022	0,3452	68,74	0,4925	0,5245	106,50
50	0,9804	3,2023	326,63	0,4965	0,3587	72,25	0,4707	0,5288	112,34

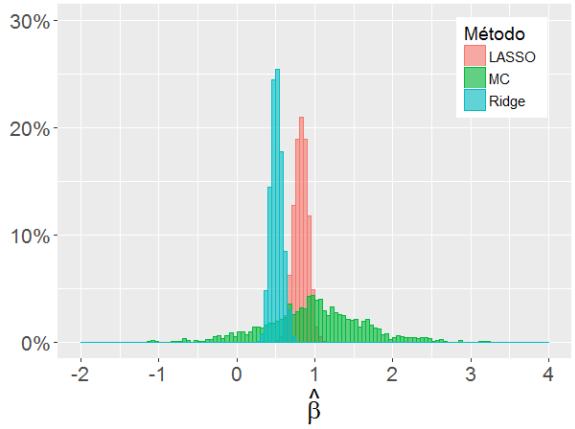
Respecto de la propiedad de insesgamiento, la Tabla 4.1 corrobora que los estimadores MC tienen promedios muy cercanos a 1, el verdadero valor del parámetro, mientras que los estimadores *Ridge* son sesgados, con promedios cercanos a 0,50 para todo k. Los estimadores

LASSO también son sesgados, pero el sesgo aumenta con k. Cuando k es pequeño, los promedios de estos últimos son cercanos al valor 1 (alrededor de 0,75), siendo comparativamente mejores que los *Ridge* en modelos con ese grado de esparcimiento.

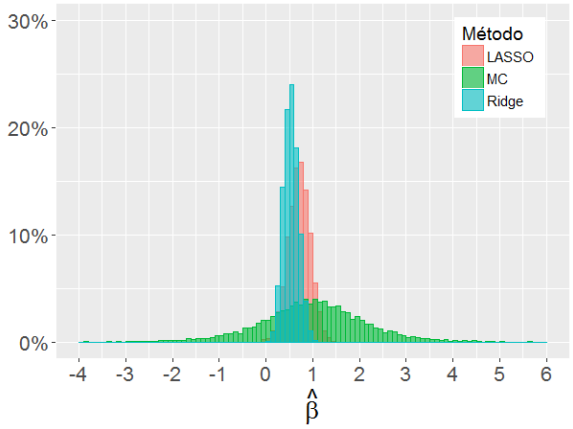
La variabilidad de los estimadores MC es mucho mayor que la de los métodos de regularización para todo k. Al comparar los coeficientes de variación de los estimadores *Ridge* y LASSO, se observa que los de LASSO son menores que los de *Ridge* en los modelos más esparcidos y que esta relación se revierte a medida que aumenta k. Para todos los métodos, los coeficientes de variación aumentan con k.

En la Figura 4.5 se comparan las distribuciones de los estimadores de $\beta = 1$ para los tres métodos de estimación, evidenciando algunas características adicionales a las señaladas a partir del análisis de la Tabla 4.1. Las distribuciones de los estimadores presentan un contorno totalmente distinto según el método aplicado. Las diferencias más notables se dan entre MC y los dos métodos de regularización. La distribución de los estimadores MC se asemeja a la distribución Normal y es mucho más dispersa. Para captar con mayor detalle las diferencias entre las regresiones *Ridge* y LASSO, la Figura 4.6 repite las distribuciones de los estimadores $\hat{\beta}^{(1)}$ omitiendo los obtenidos por MC.

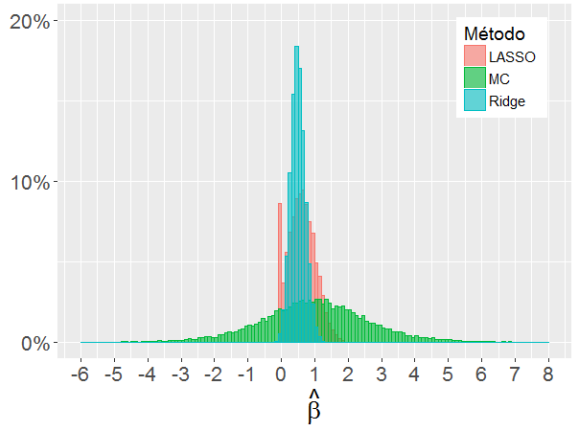
Los estimadores *Ridge* muestran distribuciones empíricas con forma campanular para todo k, evidenciando su distribución teórica Normal, con menor dispersión para k pequeños. A medida que k crece, la distribución se achata y se expande su variabilidad (Figura 4.6). Para valores de k pequeños, los estimadores LASSO también muestran una distribución de frecuencias campanular, pero para $k \geq 20$ (aproximadamente) se encuentra un gran porcentaje de estimadores nulos aún cuando el parámetro a estimar es igual a 1. Esta característica es la que empeora el sesgo evidenciado en la Tabla 4.1.



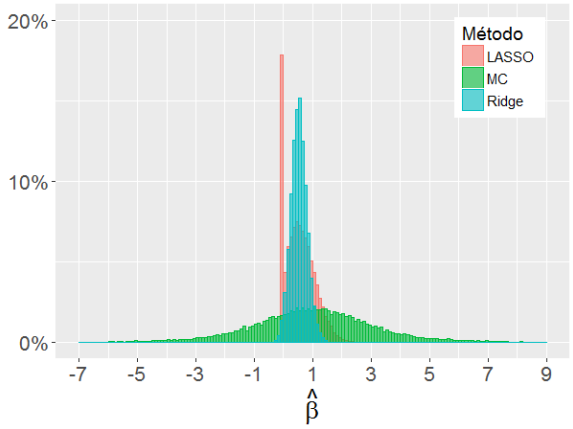
(a) $k = 2$



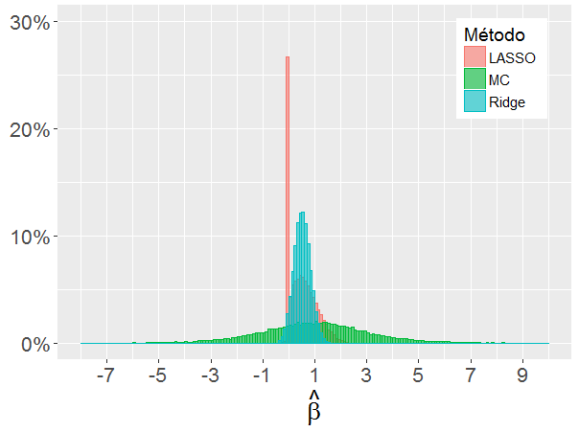
(b) $k = 10$



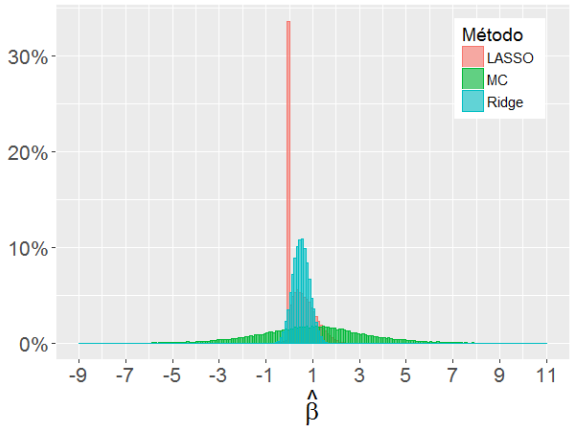
(c) $k = 20$



(d) $k = 30$

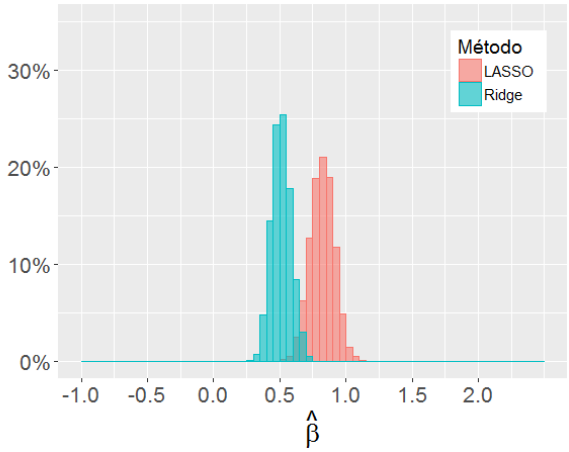


(e) $k = 40$

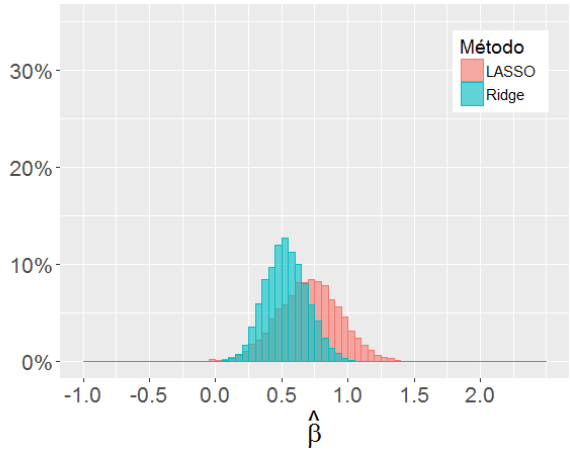


(f) $k = 50$

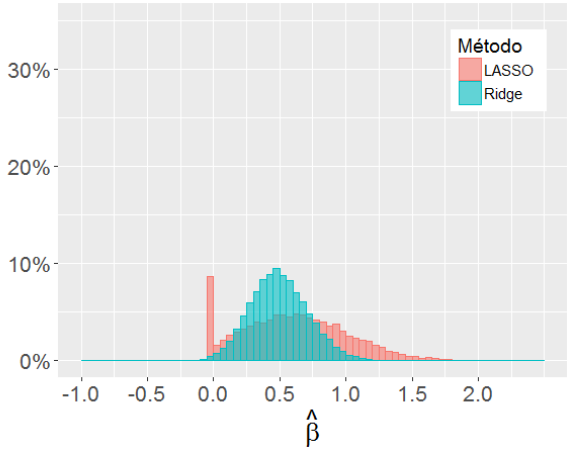
Figura 4.5: Distribución empírica de los estimadores $\hat{\beta}^{(1)}$ para variantes de k cuando $p = n = 100$. Métodos MC, *Ridge* y LASSO



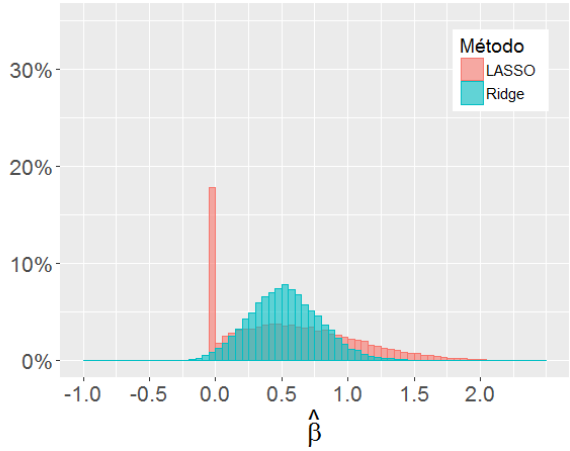
(a) $k = 2$



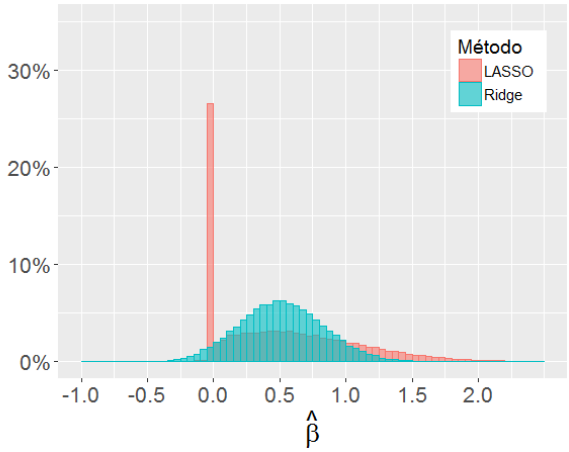
(b) $k = 10$



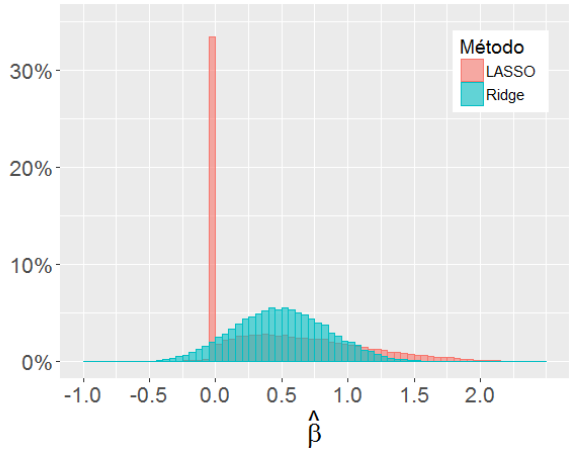
(c) $k = 20$



(d) $k = 30$



(e) $k = 40$



(f) $k = 50$

Figura 4.6: Distribución empírica de los estimadores $\hat{\beta}^{(1)}$ para variantes de k cuando $p = n = 100$. Métodos *Ridge* y LASSO

Situación 2 ($p = 2n = 200$)

En esta situación se restringe la comparación a los métodos *Ridge* y LASSO por las propiedades de los estimadores MC señaladas en capítulos anteriores. En la Tabla 4.2 se muestran los promedios, desvíos estándares y coeficientes de variación a través de las 1.000 repeticiones.

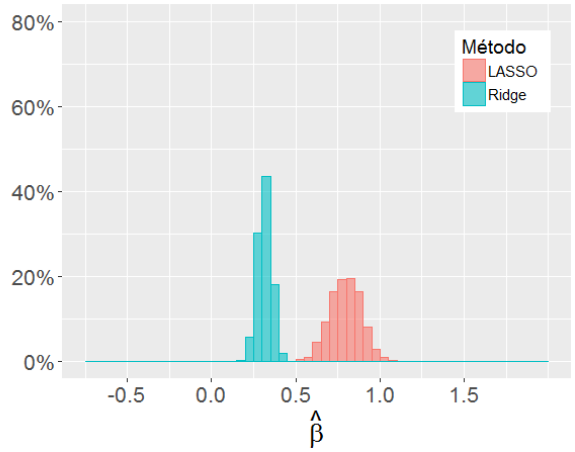
Tabla 4.2: Medidas características de las distribuciones de los estimadores $\hat{\beta}^{(1)}$ de cada método para variantes de k. Caso $p = 2n = 200$.

k	**Método**					
	Ridge			**LASSO**		
	Promedio $\overline{\hat{\beta}}^{(1)}_{Ridge\,\gamma}$	**Desvío Estándar** $S^{(1)}_{\hat{\beta}_{Ridge\,\gamma}}$	**CV(%)**	**Promedio** $\overline{\hat{\beta}}^{(1)}_{LASSO\,\gamma}$	**Desvío Estándar** $S^{(1)}_{\hat{\beta}_{LASSO\,\gamma}}$	**CV(%)**
2	0,3152	0,0420	13,32	0,7938	0,0937	11,80
4	0,2971	0,0657	22,11	0,7242	0,1469	20,28
6	0,2911	0,0712	24,46	0,6677	0,1795	26,88
8	0,3040	0,1108	36,45	0,6146	0,2296	37,36
10	0,2944	0,1093	37,13	0,5847	0,2570	43,95
15	0,3143	0,1294	41,17	0,5265	0,3409	64,75
20	0,3228	0,1500	46,47	0,4688	0,3714	79,22
25	0,3382	0,1653	48,88	0,4377	0,4211	96,21
30	0,3229	0,1725	53,42	0,3762	0,4345	115,50
35	0,3202	0,1948	60,84	0,3267	0,4338	132,78
40	0,2818	0,2016	71,54	0,2717	0,3975	146,30
45	0,2580	0,2108	81,71	0,2192	0,3676	167,70
50	0,2632	0,2189	83,17	0,2076	0,3642	175,43
60	0,2851	0,2505	87,86	0,2067	0,3815	184,57
70	0,2927	0,2690	91,90	0,2034	0,4006	196,95
80	0,2911	0,2794	95,98	0,1967	0,4117	209,30
90	0,2872	0,2925	101,85	0,1721	0,4052	235,44
100	0,2835	0,3077	108,54	0,1625	0,4018	247,26

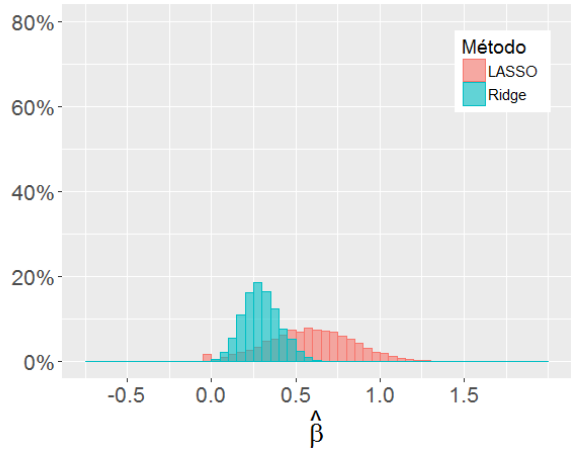
Al igual que en la Situación 1, los estimadores *Ridge* son sesgados (Tabla 4.2). Sin embargo, en este caso el sesgo es mayor, dado que los promedios son cercanos a 0,30 para todo k. Por otro lado, el sesgo de los estimadores LASSO nuevamente aumenta con k, observando sesgos mayores que en la Situación 1, sobre todo para valores de $k \geq 20$ aproximadamente.

Nuevamente, la variabilidad de los estimadores *Ridge* y LASSO aumenta con k, observando que los coeficientes de variación LASSO son menores que los *Ridge* cuando k es pequeño y que esta relación se invierte cuando k aumenta.

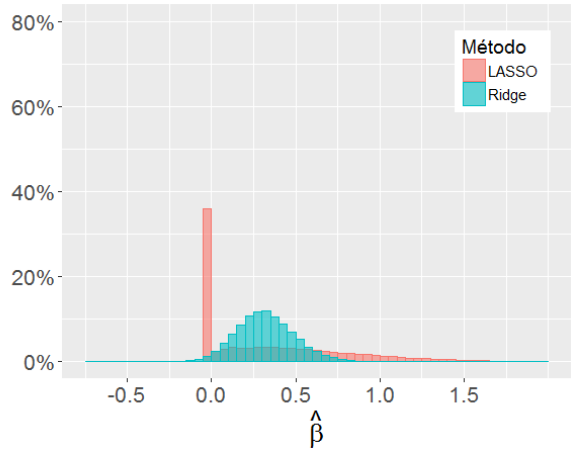
Los coeficientes de variación LASSO de la Situación 2 son mucho mayores que los de la Situación 1 (Tabla 4.1), sobre todo a partir de $k \geq 30$.



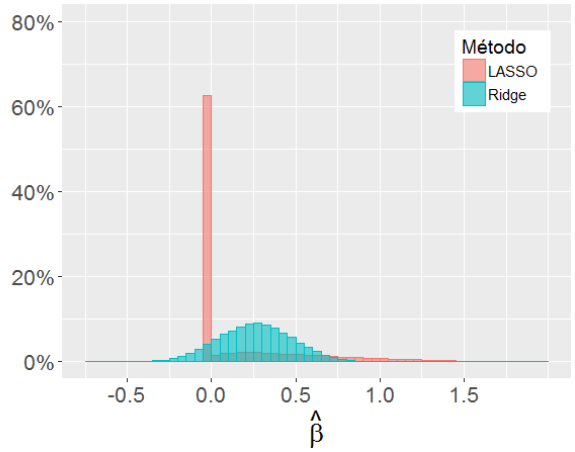
(a) $k = 2$



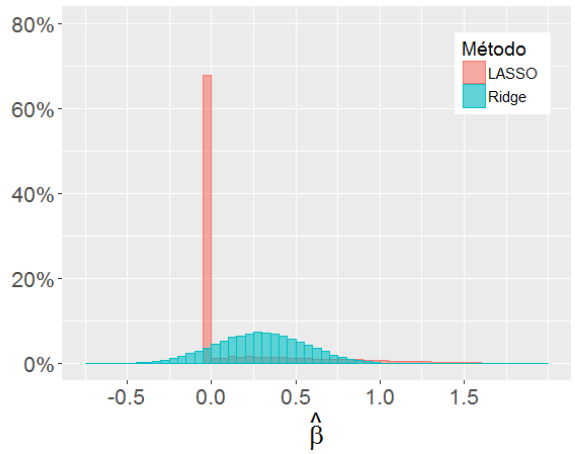
(b) $k = 10$



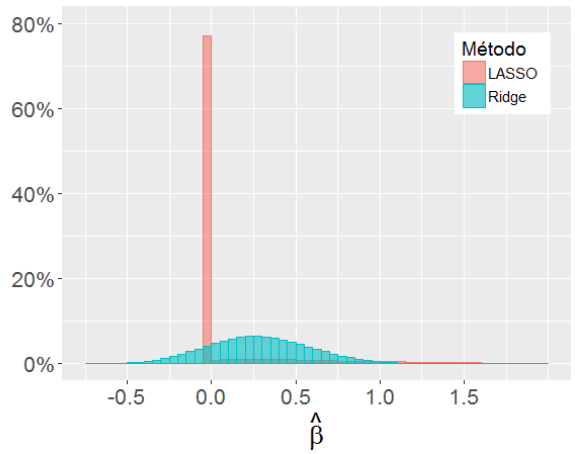
(c) $k = 30$



(d) $k = 50$



(e) $k = 70$



(f) $k = 100$

Figura 4.7: Distribución empírica de los estimadores $\hat{\beta}^{(1)}$ para variantes de k cuando $p = 2n = 200$

En la Figura 4.7 se comparan las distribuciones de los estimadores $\hat{\beta}^{(1)}$ de los métodos *Ridge* y LASSO. En esta figura vuelve a evidenciarse el comportamiento campanular de las distribuciones de los estimadores *Ridge* para todo k. Este mismo comportamiento se observa para LASSO cuando k es pequeño, pero a medida que k aumenta, también lo hace la proporción de estimadores nulos, hecho que empeora el desempeño de este método.

Situación 3 ($p = 4n = 400$)

Tabla 4.3: Medidas características de las distribuciones de los estimadores $\hat{\beta}^{(1)}$ de cada método para variantes de k. Caso $p = 4n = 400$.

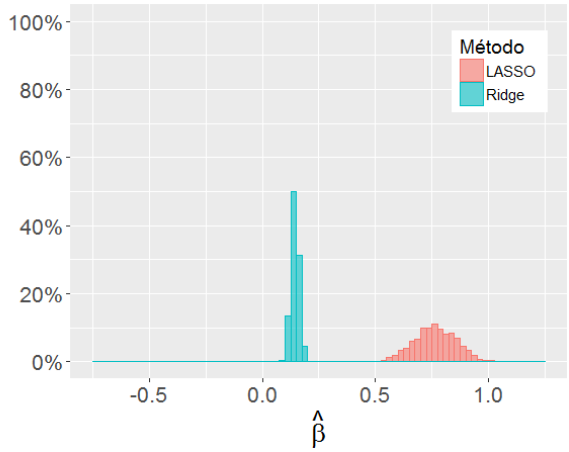
k	**Método**					
	Ridge			**LASSO**		
	Promedio $\overline{\hat{\beta}}^{(1)}_{Ridge\,\gamma}$	**Desvío Estándar** $S^{(1)}_{\hat{\beta}_{Ridge\,\gamma}}$	**CV(%)**	**Promedio** $\overline{\hat{\beta}}^{(1)}_{LASSO\,\gamma}$	**Desvío Estándar** $S^{(1)}_{\hat{\beta}_{LASSO\,\gamma}}$	**CV(%)**
2	0,1444	0,0181	12,53	0,7641	0,0951	12,45
5	0,1516	0,0355	23,42	0,6402	0,1706	26,65
10	0,1497	0,0586	39,14	0,4644	0,2688	57,88
20	0,1819	0,0874	48,05	0,3109	0,3606	115,99
30	0,1651	0,0950	57,54	0,1871	0,3336	178,30
40	0,1394	0,1080	77,47	0,1042	0,2472	237,24
50	0,1300	0,1143	87,92	0,0732	0,2076	283,61
75	0,1479	0,1420	96,01	0,0643	0,2155	335,15
100	0,1603	0,1796	112,04	0,0503	0,2137	424,85
150	0,1730	0,2277	131,62	0,0405	0,2082	514,07
200	0,1764	0,2726	154,54	0,0419	0,2223	530,55

Nuevamente, los estimadores *Ridge* son sesgados, en este caso con promedios cercanos a 0,15 para todo k (Tabla 4.3). La variabilidad de estos estimadores aumenta con k, siendo mayor que la observada en las Situaciones 1 y 2.

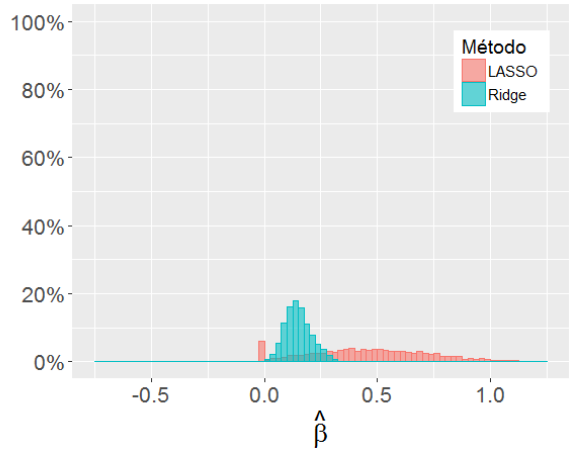
En la Situación 3, el sesgo de los estimadores LASSO es mayor que en las situaciones anteriores, y aumenta con mayor rapidez. A partir de $k \geq 50$ aproximadamente, se estabiliza. Al igual que lo observado para los estimadores *Ridge*, los coeficientes de variación crecen con k, observándose valores mucho mayores que en los casos $p = n = 100$ y $p = 2n = 200$.

A partir de $k = 5$ aproximadamente, los coeficientes de variación de los estimadores *Ridge*

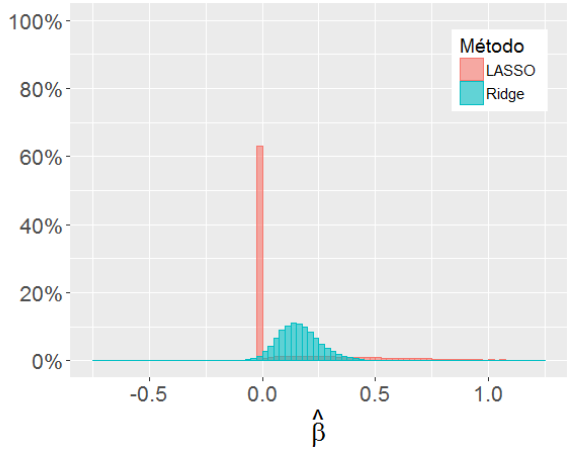
son menores que los de LASSO.



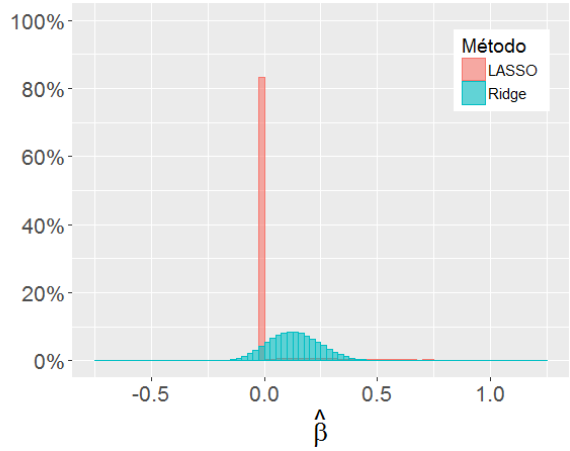
(a) $k = 2$



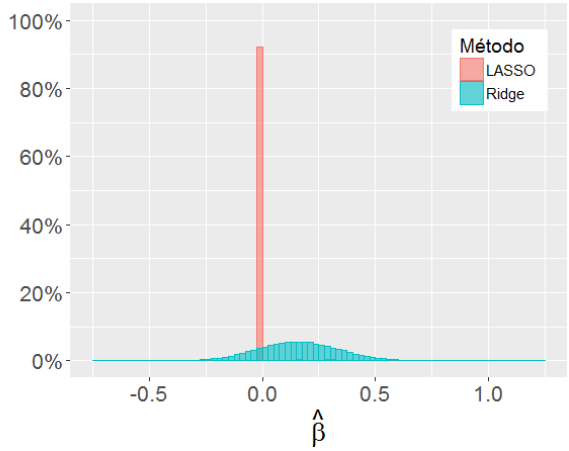
(b) $k = 10$



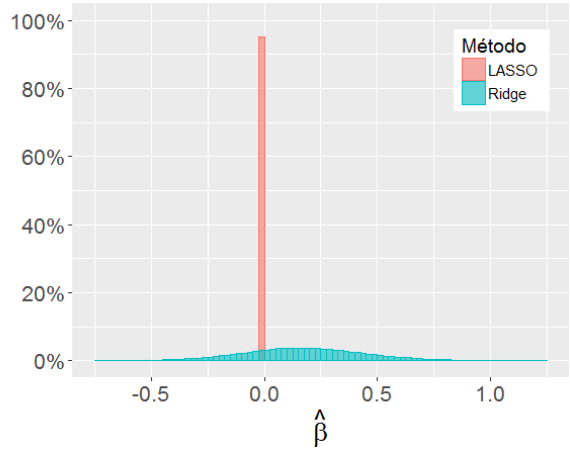
(c) $k = 30$



(d) $k = 50$



(e) $k = 100$



(f) $k = 200$

Figura 4.8: Distribución empírica de los estimadores $\hat{\beta}^{(1)}$ para variantes de k cuando $p = 4n = 400$

En la Figura 4.8, donde se comparan las distribuciones empíricas de los estimadores de

$\beta = 1$ de los métodos *Ridge* y LASSO, se observan comportamientos idénticos a los observados en las Situaciones 1 y 2. Los estimadores *Ridge* evidencian su conocida distribución Normal, y los LASSO presentan forma campanular hasta el momento en el que comienza a aumentar la proporción de estimaciones nulas. A partir de $k = 30$ aproximadamente, la mayoría de las estimaciones LASSO son nulas, y este es el motivo por el cual disminuyen los desvíos estándares para este método en la Tabla 4.3.

4.4. Propiedades Distribucionales de Estimadores $\hat{\beta}^{(0)}$

En esta Sección se comparan los promedios y desvíos estándares de los estimadores de los parámetros $\beta = 0$ para cada método en cada escenario, agrupando escenarios por situación. A su vez, debido a la regularidad de los resultados obtenidos, se muestran gráficamente las distribuciones empíricas de los estimadores $\hat{\beta}^{(0)}$ para cada método sólo en algunos escenarios seleccionados. Los restantes pueden consultarse en el Anexo V.

Situación 1 $(p = n = 100)$

Con respecto a la propiedad de insesgamiento, la Tabla 4.4 corrobora que los estimadores MC tienen promedios muy cercanos a 0 para todo k. Lo mismo se observa para los estimadores de los métodos de regularización.

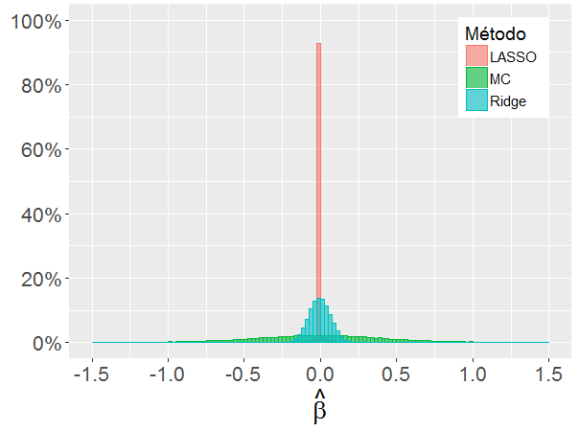
Las diferencias entre los métodos se observan en la variabilidad de sus estimaciones. Los desvíos estándares de los estimadores MC son severamente mayores que los de los métodos de regularización. Al comparar los desvíos de los estimadores *Ridge* y LASSO, se observa que los de LASSO son siempre menores que los de *Ridge*, y que esta diferencia disminuye a medida que k aumenta.

Tabla 4.4: Medidas características de las distribuciones de los estimadores $\hat{\beta}^{(0)}$ de cada método para variantes de k. Caso $p = n = 100$.

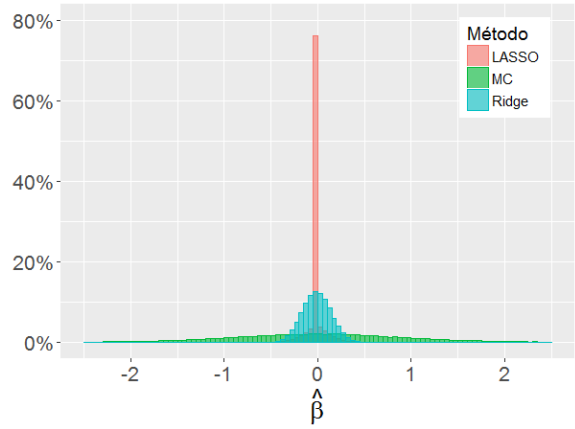
k	**Método**					
	MC		**Ridge**		**LASSO**	
	Promedio $\overline{\hat{\beta}}^{(0)}_{MC\,\gamma}$	**Desvío Estándar** $S^{(0)}_{\hat{\beta}_{MC\,\gamma}}$	**Promedio** $\overline{\hat{\beta}}^{(0)}_{Ridge\,\gamma}$	**Desvío Estándar** $S^{(0)}_{\hat{\beta}_{Ridge\,\gamma}}$	**Promedio** $\overline{\hat{\beta}}^{(0)}_{LASSO\,\gamma}$	**Desvío Estándar** $S^{(0)}_{\hat{\beta}_{LASSO\,\gamma}}$
2	0,0008	0,6161	-0,0063	0,0701	-0,0005	0,0165
4	0,0004	0,9163	-0,0062	0,0997	-0,0006	0,0337
6	-0,0003	1,0949	-0,0036	0,1230	-0,0004	0,0516
8	-0,0002	1,2753	-0,0098	0,1410	-0,0026	0,0729
10	0,0007	1,4148	-0,0133	0,1562	-0,0034	0,0883
15	0,0037	1,8268	-0,0143	0,1909	-0,0061	0,1330
20	0,0048	2,1981	-0,0207	0,2186	-0,0109	0,1743
25	0,0027	2,4424	-0,0049	0,2530	-0,0038	0,2126
30	0,0035	2,6054	-0,0140	0,2757	-0,0119	0,2409
35	0,0054	2,9075	-0,0052	0,2957	-0,0069	0,2637
40	0,0119	2,9093	0,0027	0,3188	0,0003	0,2852
45	0,0118	2,9018	0,0259	0,3428	0,0169	0,3219
50	0,0135	3,1190	0,0180	0,3600	0,0076	0,3356

En la Figura 4.9, donde se comparan las distribuciones de los estimadores de los parámetros $\beta = 0$ obtenidos por el ajuste mínimo-cuadrático, regresión *Ridge* y regresión LASSO, se evidencia lo explicitado a partir de los desvíos estándares (Tabla 4.4). La distribución de los estimadores MC es mucho más dispersa que la de los métodos de regularización. En esta figura se observa que los estimadores *Ridge* muestran distribuciones campanulares para todo k.

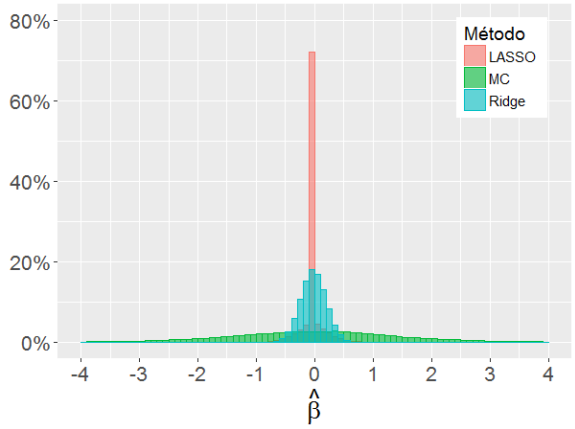
Dada la gran variabilidad de los estimadores MC en relación a la de los métodos de regularización, la Figura 4.10 repite estas distribuciones sólo para los estimadores de los métodos *Ridge* y LASSO. En ella se observa mejor cómo la variabilidad de ambos métodos aumenta con k. Una consecuencia de esto es la disminución de la proporción de estimadores LASSO que son nulos a medida que k aumenta.



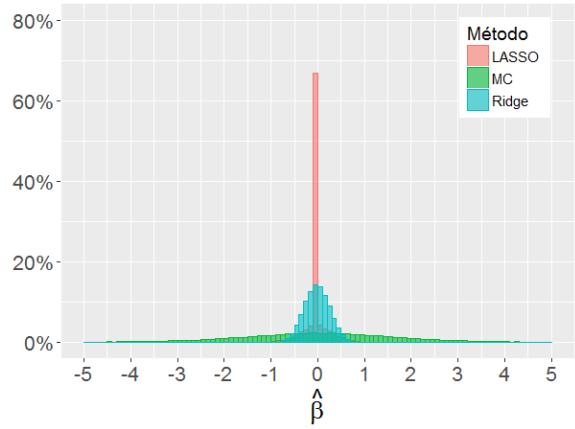
(a) $k = 2$



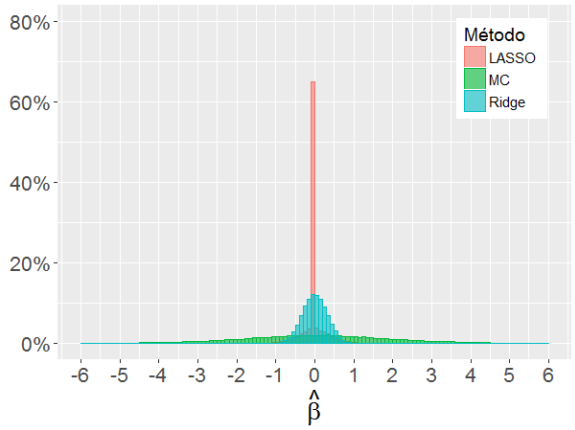
(b) $k = 10$



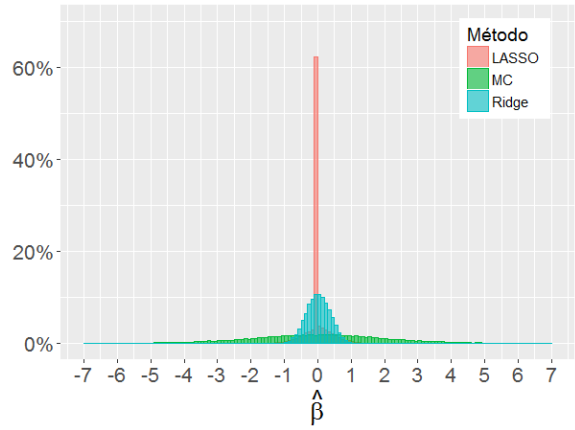
(c) $k = 20$



(d) $k = 30$

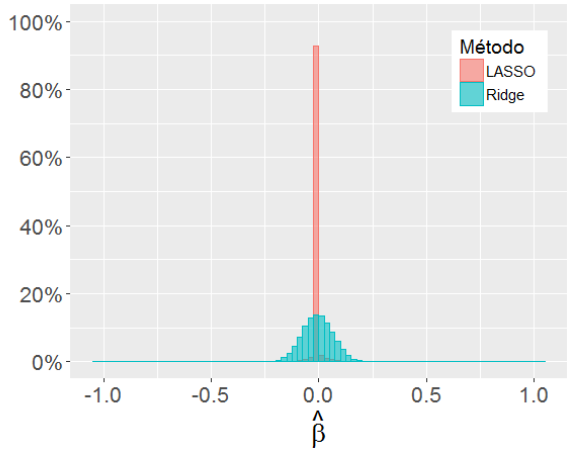


(e) $k = 40$

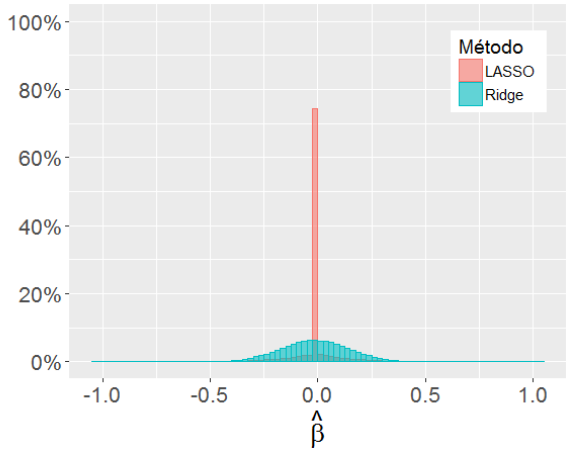


(f) $k = 50$

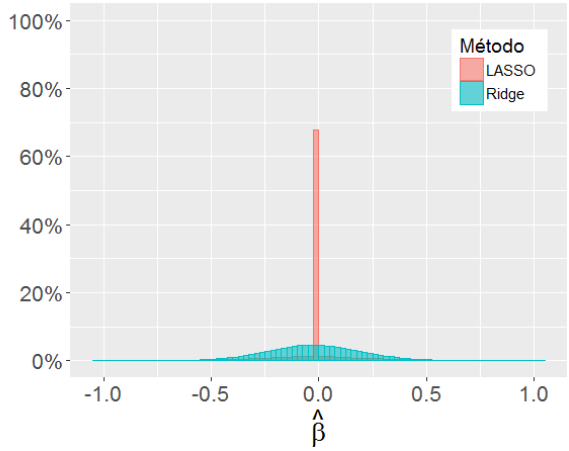
Figura 4.9: Distribución empírica de los estimadores $\hat{\beta}^{(0)}$ para variantes de k cuando $p = n = 100$. Métodos MC, *Ridge* y LASSO



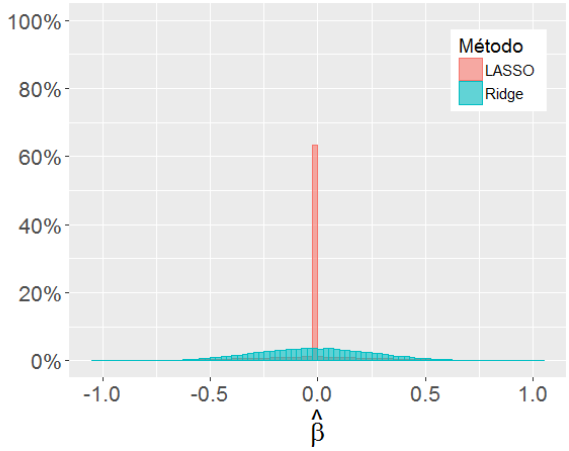
(a) $k = 2$



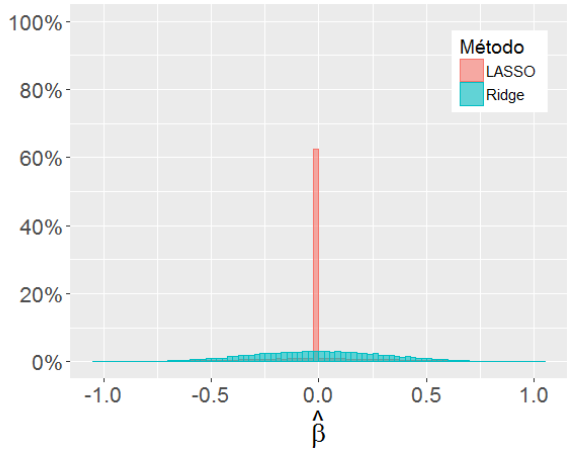
(b) $k = 10$



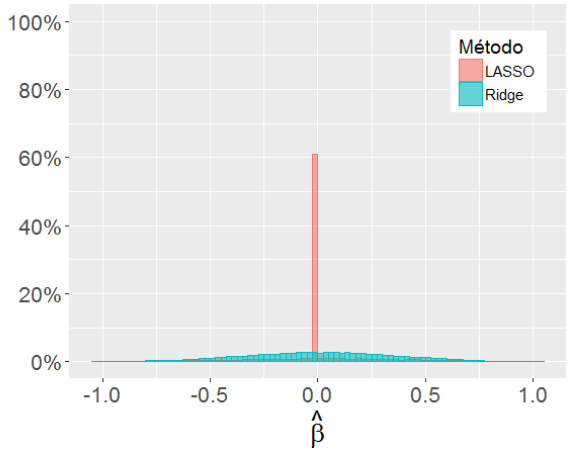
(c) $k = 20$



(d) $k = 30$



(e) $k = 40$



(f) $k = 50$

Figura 4.10: Distribución empírica de los estimadores $\hat{\beta}^{(0)}$ para variantes de k cuando $p = n = 100$. Métodos *Ridge* y LASSO

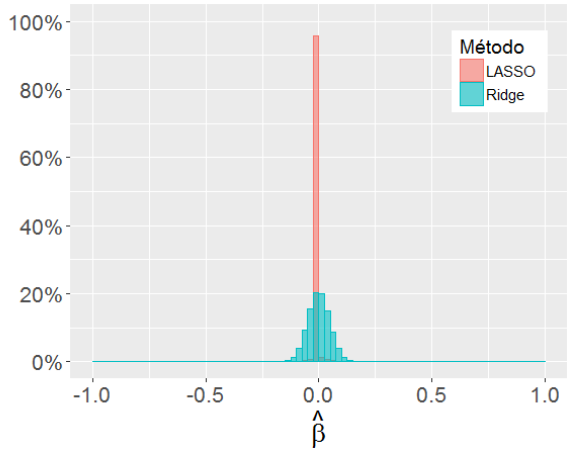
Situación 2 ($p = 2n = 200$)

Tabla 4.5: Medidas características de las distribuciones de los estimadores $\hat{\beta}^{(0)}$ de cada método para variantes de k. Caso $p = 2n = 200$.

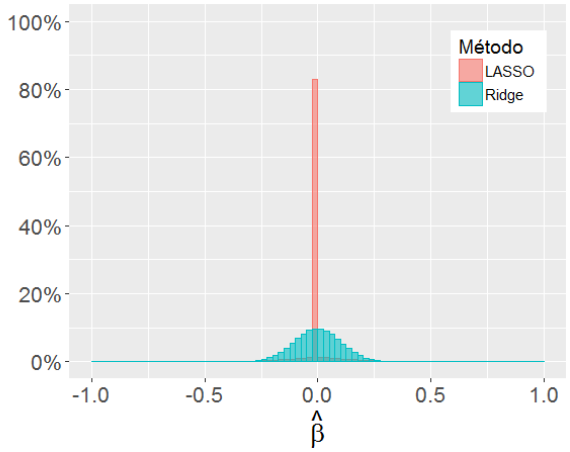
k	**Método**			
	Ridge		**LASSO**	
	Promedio $\overline{\hat{\beta}}^{(0)}_{Ridge\,\gamma}$	**Desvío Estándar** $S^{(0)}_{\hat{\beta}_{Ridge\,\gamma}}$	**Promedio** $\overline{\hat{\beta}}^{(0)}_{LASSO\,\gamma}$	**Desvío Estándar** $S^{(0)}_{\hat{\beta}_{LASSO\,\gamma}}$
2	-0,0011	0,0471	-0,0001	0,0121
4	0,0001	0,0652	0,0000	0,0248
6	0,0018	0,0801	0,0004	0,0415
8	0,0023	0,0931	0,0009	0,0622
10	-0,0010	0,1026	0,0005	0,0734
15	-0,0021	0,1279	0,0000	0,1041
20	0,0023	0,1482	0,0014	0,1264
25	0,0066	0,1719	0,0041	0,1549
30	0,0036	0,1868	-0,0008	0,1612
35	-0,0004	0,1975	-0,0035	0,1575
40	0,0051	0,2012	-0,0037	0,1680
45	0,0124	0,2071	-0,0012	0,1775
50	0,0059	0,2198	-0,0055	0,1715
60	0,0136	0,2473	-0,0011	0,1872
70	0,0109	0,2719	-0,0066	0,2055
80	0,0127	0,2894	-0,0041	0,2239
90	0,0158	0,3087	-0,0033	0,2341
100	0,0225	0,3260	0,0018	0,2451

Al igual que en el caso $p = n = 100$, los promedios de los estimadores *Ridge* y LASSO son cercanos a 0 (Tabla 4.5). Los desvíos estándares de los estimadores *Ridge* son siempre mayores que los de LASSO, y esta diferencia disminuye a medida que k aumenta.

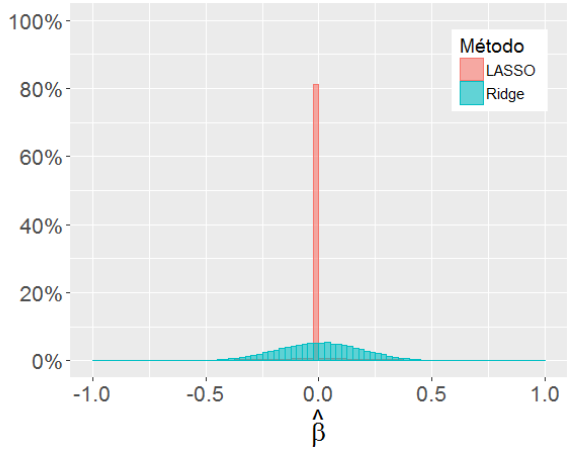
En la Figura 4.11 se comparan las distribuciones de los estimadores $\hat{\beta}^{(0)}$ de los métodos *Ridge* y LASSO. En ella se observa el comportamiento campanular de la distribución de los estimadores *Ridge*, el cual se hace menos visible para valores de k grandes debido al aumento de su variabilidad y por haber mantenido la escala para dibujar simultáneamente las distribuciones para ambos métodos. Con respecto a estos últimos, se observa que el porcentaje de estimaciones nulas disminuye con k hasta $k = 30$ aproximadamente, y a partir de ese valor el mismo aumenta levemente con k. De todos modos, esta proporción se mantiene alrededor del 85 % en todo momento.



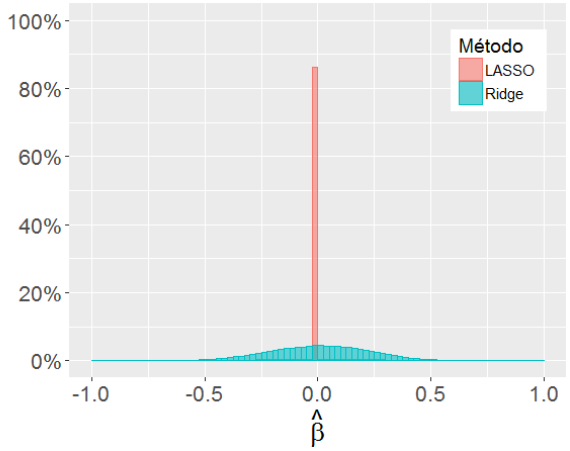
(a) $k = 2$



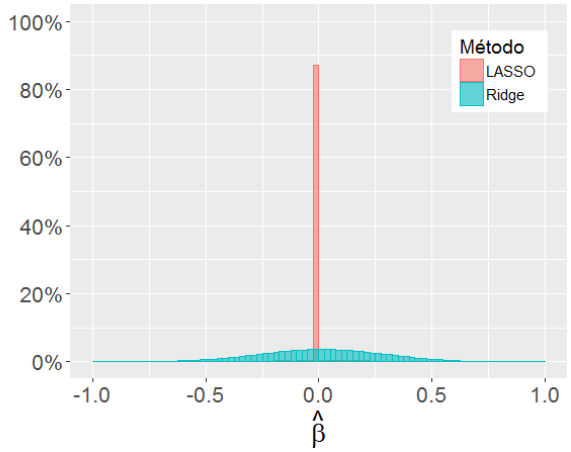
(b) $k = 10$



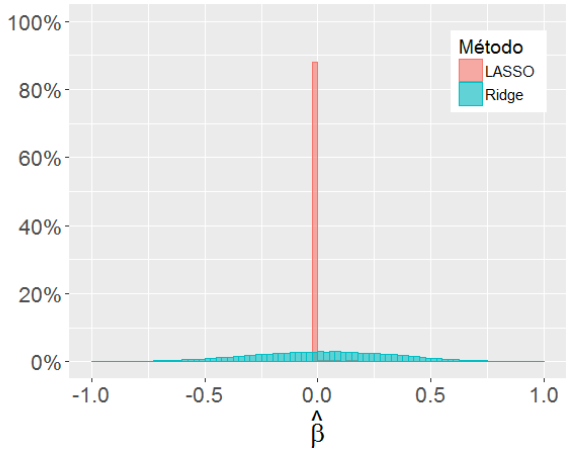
(c) $k = 30$



(d) $k = 50$



(e) $k = 70$



(f) $k = 100$

Figura 4.11: Distribución empírica de los estimadores $\hat{\beta}^{(0)}$ para variantes de k cuando $p = 2n = 200$

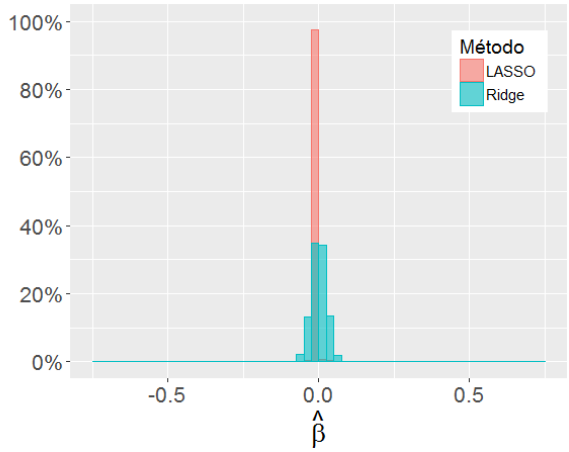
Situación 3 ($p = 4n = 400$)

Tabla 4.6: Promedio y desvío estándar de las distribuciones de los estimadores de parámetros $\beta = 0$ para cada método en cada escenario de la Situación 3

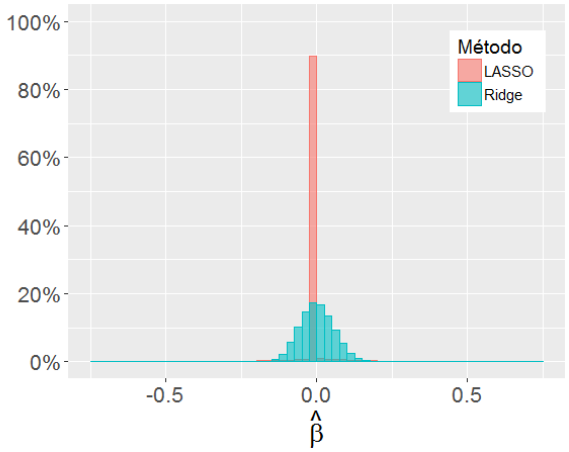
k	**Método**			
	Ridge		**LASSO**	
	Promedio $\bar{\hat{\beta}}^{(0)}_{Ridge\,\gamma}$	**Desvío Estándar** $S^{(0)}_{\hat{\beta}_{Ridge\,\gamma}}$	**Promedio** $\bar{\hat{\beta}}^{(0)}_{LASSO\,\gamma}$	**Desvío Estándar** $S^{(0)}_{\hat{\beta}_{LASSO\,\gamma}}$
2	-0,0001	0,0247	0,0001	0,0087
5	0,0019	0,0402	0,0004	0,0251
10	0,0001	0,0564	0,0011	0,0567
20	0,0012	0,0871	0,0024	0,0806
30	0,0010	0,1030	0,0002	0,0832
40	0,0002	0,1091	-0,0015	0,0876
50	-0,0006	0,1184	-0,0018	0,0854
75	-0,0014	0,1543	-0,0028	0,0973
100	-0,0022	0,1841	-0,0027	0,0898
150	-0,0040	0,2328	-0,0011	0,0946
200	-0,0104	0,2668	-0,0005	0,0926

Nuevamente, los promedios de los estimadores *Ridge* y LASSO son cercanos a 0 para todo valor de k (Tabla 4.3). Los desvíos estándares de los estimadores *Ridge* aumentan con k, mientras que los de LASSO aumentan hasta $k = 20$ aproximadamente, manteniéndose relativamente constantes a partir de dicho valor. Se observa que la variabilidad de los estimadores LASSO es siempre menor que la de los *Ridge*, repitiendo el comportamiento de las Situaciones 1 y 2.

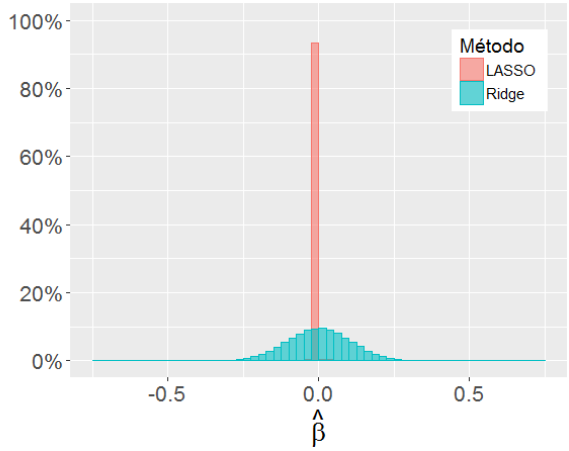
En la Figura 4.12, donde se presentan las distribuciones de los estimadores $\hat{\beta}^{(0)}$ para cada método, se observa que los estimadores *Ridge* se comportan igual que en el caso $p = 2n = 200$. En la Situación 3, la proporción de estimaciones LASSO nulas disminuye hasta $k = 10$ aproximadamente, y luego aumenta con k, manteniendo proporciones cercanas a 1, siendo este el motivo por el cual la variabilidad de las estimaciones LASSO se estabiliza para valores de k grandes.



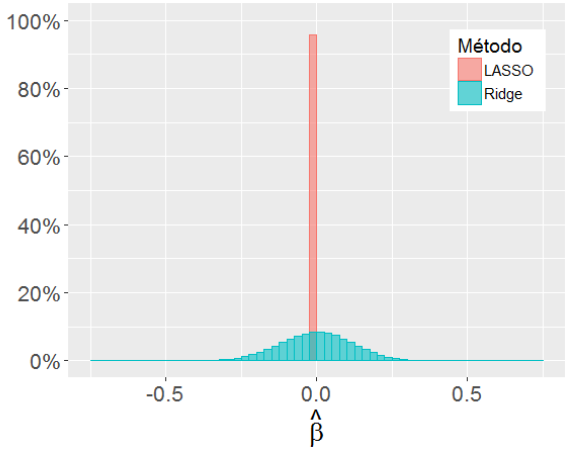
(a) $k = 2$



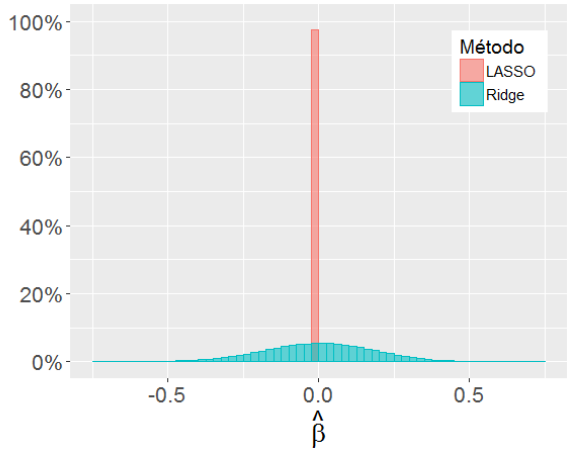
(b) $k = 10$



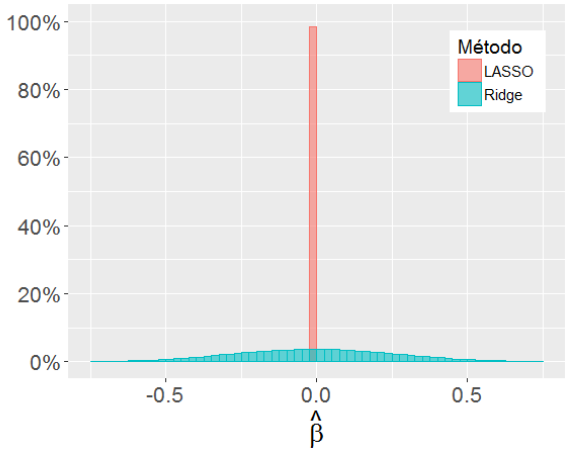
(c) $k = 30$



(d) $k = 50$



(e) $k = 100$



(f) $k = 200$

Figura 4.12: Distribución empírica de los estimadores $\hat{\beta}^{(0)}$ para variantes de k cuando $p = 4n = 400$

4.5. Medidas Específicas de Estimadores LASSO

En esta Sección se estudia la capacidad de los estimadores LASSO de obtener estimaciones nulas, tanto para la situación no deseable donde el parámetro a estimar es $\beta = 1$, como para la situación deseable donde $\beta = 0$. Los resultados se presentan para las variantes de k, separados por situación.

Situación 1 ($p = n = 100$)

Tabla 4.7: Medidas específicas calculadas para los estimadores LASSO. Caso $p = n = 100$.

k	$\hat{\beta}^{(0)} = 0$ (%)	$\hat{\beta}^{(1)} = 0$ (%)
2	90,59	0,00
4	85,03	0,00
6	79,34	0,00
8	74,23	0,01
10	71,95	0,27
15	68,20	2,87
20	66,25	8,65
25	62,83	12,33
30	61,86	17,76
35	62,66	23,79
40	60,76	26,36
45	56,61	29,77
50	58,57	33,04

El porcentaje de $\hat{\beta}^{(0)}$ que son nulos disminuye a medida que k aumenta (Tabla 4.7). Esto está relacionado con el aumento en la variabilidad de las estimaciones presente en los resultados de la Tabla 4.4. Por otro lado, el porcentaje de $\hat{\beta}^{(1)}$ que son nulos aumenta con k, lo cual muestra un mal desempeño de los estimadores LASSO cuando k es grande.

Situación 2 ($p = 2n = 200$)

Tabla 4.8: Medidas específicas calculadas para los estimadores LASSO. Caso $p = 2n = 200$.

k	$\hat{\beta}^{(0)} = 0$ (%)	$\hat{\beta}^{(1)} = 0$ (%)
2	94,45	0,00
4	91,10	0,00
6	86,83	0,02
8	82,49	0,73
10	81,54	1,80
15	80,17	8,66
20	80,55	19,21
25	78,70	25,43
30	80,30	35,77
35	83,40	45,14
40	83,39	51,29
45	83,46	59,00
50	85,32	62,30
60	85,87	64,45
70	85,96	67,33
80	86,35	70,01
90	86,81	75,06
100	86,22	76,47

En este caso, el porcentaje de parámetros nulos estimados por cero por LASSO disminuye hasta $k = 25$ y luego aumenta con k, manteniéndose siempre en valores altos (4.8). Por otro lado, el porcentaje de parámetros $\beta = 1$ cuyas estimaciones son nulas aumenta con k, observando porcentajes mucho mayores que en el caso $p = n = 100$.

Situación 3 ($p = 4n = 400$)

Tabla 4.9: Medidas específicas calculadas para los estimadores LASSO. Caso $p = 4n = 400$.

k	$\hat{\beta}^{(0)} = 0$ (%)	$\hat{\beta}^{(1)} = 0$ (%)
2	96,90	0,00
5	93,23	0,02
10	89,04	6,14
20	91,07	37,58
30	92,92	61,83
40	94,59	76,43
50	95,37	83,01
75	95,78	87,15
100	96,90	91,33
150	97,65	94,14
200	97,92	94,20

Nuevamente, se observa que el porcentaje de $\hat{\beta}^{(0)} = 0$ disminuye con k para valores de k pequeños, en este caso hasta $k = 10$ aproximadamente, y partir de dicho valor ese porcentaje aumenta con k (Tabla 4.9). Se observan porcentajes muy altos en todo momento, mayores que los de la Tabla 4.8. Como en las Situaciones 1 y 2, el porcentaje de $\hat{\beta}^{(1)} = 0$ aumenta con k, observando en este caso porcentajes mucho más elevados que en las otras dos situaciones. A partir de $k = 50$ aproximadamente, casi la totalidad de las estimaciones LASSO son nulas, tanto aquellas que estiman parámetros nulos como las que estiman parámetros no nulos.

Capítulo 5

Conclusiones

El trabajo desarrollado se aboca a una problemática frecuente en el contexto actual del análisis de datos, estudiando modelos estimados a partir de bases de datos de grandes dimensiones. El uso de modelos en estadística es una herramienta utilizada con mucha frecuencia por la riqueza que brinda al analizar relaciones entre variables y la posibilidad de predecir ciertas variables de interés bajo condiciones particulares. Clásicamente, cuando se plantean modelos lineales, la estimación de sus parámetros se hace bajo el método de mínimos cuadrados, el cual tiene excelentes propiedades cuando el número de observaciones es considerablemente mayor que el de parámetros a estimar. Las actuales condiciones de recolección de datos, facilitada por el uso de mecanismos automatizados, implican frecuentemente que ésta condición no se verifique y aparezca la necesidad de estimar muchos parámetros con una cantidad pequeña de observaciones. Para estas situaciones se han propuesto métodos alternativos de estimación, dos de los cuales son objeto de un análisis más profundo en esta tesina: las regresiones *Ridge* y LASSO. Estas técnicas son métodos de regularización que imponen restricciones adicionales para la estimación de los parámetros.

En este trabajo se hace una síntesis de los métodos de estimación de los parámetros de los modelos lineales tanto en el contexto clásico (mínimos cuadrados) como los métodos más

novedosos, aplicables en el contexto de grandes dimensiones (selección del mejor subconjunto, regresión *Ridge*, regresión LASSO y red elástica). Sólo se considera la situación de variables predictivas y de respuesta continuas, señalando los criterios de optimización que requiere cada método, las propiedades de los estimadores, cómo obtener predicciones y su variabilidad, y algunos métodos computacionales imprescindibles para la aplicación de la regresión LASSO.

En particular, mínimos cuadrados no tiene solución única cuando el número de variables explicativas es mayor que el número de observaciones, motivo por el cual es necesario recurrir a métodos como los de regularización. En estos últimos, la solución depende de la elección de un parámetro de suavizado, dando lugar a un camino de soluciones. De todas las soluciones calculadas para algunas variantes de este parámetro previamente establecidas, se elige la mejor de acuerdo a algún criterio, como por ejemplo, minimización del ECM, recomendándose el uso de validación cruzada. Previo a la aplicación de los métodos de regularización resulta conveniente estandarizar los predictores, para evitar que la elección del parámetro de suavizado se vea influenciada por la escala de los mismos. El estimador *Ridge* puede ser calculado unívocamente y siempre tiene una forma cerrada para su cálculo. LASSO admite solución en forma cerrada únicamente cuando la matriz de predictores es ortogonal. Cuando este no es el caso, obtener el estimador LASSO constituye un problema de programación cuadrático, cuya solución puede ser aproximada eficientemente. En este trabajo también se mencionan características particulares de los problemas de optimización planteados por los métodos *Ridge* y LASSO y se hace una breve descripción de un software libre que puede ser utilizado para implementar la estimación del método LASSO, el cual es intensivo computacionalmente.

En esta tesina se plantea un estudio por simulación diseñado para mostrar las propiedades de estos métodos frente a distintos contextos de volumen de información disponible respecto de la cantidad de parámetros a estimar. Se definen tres situaciones distintas, una donde el número de observaciones es igual al número de parámetros y dos en las cuales hay el doble

y el cuádruple de parámetros respecto del número de observaciones. En cada una de las situaciones se construyen distintos modelos variando el grado de esparcimiento (k) de los mismos, siendo k el número de parámetros no nulos del modelo. Sin pérdida de generalidad, los parámetros no nulos se fijan en uno. A fin de establecer una comparación entre los métodos se identificaron medidas globales de eficiencia, a saber: ECM de los modelos, variabilidad y sesgo de los estimadores de los parámetros separando en dos casos, nulos y no nulos, forma de su distribución empírica y capacidad del método LASSO para identificar parámetros nulos y no nulos del modelo. Esta última medida de eficiencia se introduce debido a la particularidad que presentan los modelos ajustados por este método en tablas de grandes dimensiones, donde es frecuente que existan muchas variables explicativas que no tengan una influencia significativa sobre la respuesta.

Los resultados de estas simulaciones muestran que la capacidad predictiva de los estimadores mínimo-cuadráticos, medida en términos del ECM, resulta peor que la de los métodos de regularización. Si bien el ajuste mínimo-cuadrático provee estimadores insesgados tanto para los parámetros nulos como los no nulos, la variabilidad de los mismos es mucho mayor que la de las regresiones penalizadas.

En todas las situaciones, al comparar los $\overline{ECM}$ de los métodos de regularización, se observó que la regresión LASSO tiene mejor desempeño que *Ridge* en los modelos más esparcidos y que esta relación se revierte en modelos densos. Resulta llamativo que en las tres situaciones, pese a que el número de predictores es diferente en cada caso, el cruce de las curvas de estos dos métodos se da en valores cercanos a $k = 25$.

Con respecto a la estimación de parámetros no nulos, los estimadores *Ridge* resultan sesgados en todas las situaciones, siendo el sesgo constante para los distintos niveles de esparcimiento dentro de cada situación. Se observan sesgos mayores frente a mayor cantidad de variables explicativas. El desempeño de LASSO en este sentido es similar, es decir, se

observan en general resultados peores en las situaciones con mayor número de variables explicativas. Sin embargo, dentro de cada caso, LASSO presenta menor sesgo que *Ridge* en los modelos más esparcidos y el mismo aumenta con k, llegando a ser mayor que el de *Ridge* en los modelos más densos. La variabilidad de los estimadores LASSO es menor para valores de k pequeños, y esta relación se revierte al aumentar k. En las tres situaciones este cambio se produce cerca del valor $k = 5$.

Con respecto a la estimación de parámetros nulos, los estimadores de los métodos de regularización resultan insesgados en todas las situaciones. La variabilidad de los estimadores LASSO es siempre menor que la de los estimadores *Ridge*, y dentro de cada situación esta diferencia disminuye a medida que aumenta k, a excepción de la Situación 3, donde la variabilidad de los estimadores LASSO se estabiliza.

La habilidad de LASSO para estimar con cero a parámetros nulos es muy buena en todas las situaciones, presentando mejor desempeño en aquellas donde el número de variables explicativas es mayor. Sin embargo, también tiende a estimar con cero muchos parámetros no nulos a medida que aumenta la cantidad de parámetros significativos. Este puede ser uno de los motivos por los cuales el desempeño de LASSO empeora a medida que aumenta k.

En general, todos los métodos empeoran su desempeño cuando el número de parámetros no nulos es grande. Esto se relaciona con el principio "*bet on sparsity*" (Hastie et al., 2009), que propone elegir aquel procedimiento que funcione bien en problemas esparcidos, ya que ningún método funciona bien en problemas densos.

A modo de recomendación final, y a partir de las propiedades estudiadas, puede enunciarse que, como con LASSO se obtienen mejores resultados en problemas esparcidos, sería preferible en lugar de *Ridge* en contextos de grandes dimensiones de datos, mientras que mínimos cuadrados no es una alternativa admisible en estos escenarios.

Como líneas de investigación para futuros trabajos en temáticas vinculadas a la considerada en esta tesina, pero que no fueron tratadas, se pueden mencionar:

- El uso de variables explicativas discretas
- El uso de variables explicativas cualitativas
- Empleo de distintas versiones de la regresión LASSO, como *grouped* LASSO, *relaxed* LASSO, *adaptive* LASSO y *bayesian* LASSO

También sería de interés profundizar en el estudio empírico y teórico de la forma distribucional de los estimadores LASSO dadas las características que se han observado en la presente tesina.

Referencias

Beale, E., Kendall, M., and Mann, D. (1967). The discarding of variables in multivariate analysis. *Biometrika*, 54(3-4):357–366.

Bien, J. (2016). The simulator: An engine to streamline simulations. *Submitted*.

Boyd, S. and Vandenberghe, L. (2004). *Convex optimization*. Cambridge, UK: Cambridge University Press.

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

Friedman, J., Hastie, T., Höfling, H., Tibshirani, R., et al. (2007). Pathwise coordinate optimization. *The Annals of Applied Statistics*, 1(2):302–332.

Friedman, J., Hastie, T., and Tibshirani, R. (2009). glmnet: Lasso and elastic-net regularized generalized linear models. *R package version*, 1(4).

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, 33(1):1.

Galton, F. (1886). Regression towards mediocrity in hereditary stature. *The Journal of the Anthropological Institute of Great Britain and Ireland*, 15:246–263.

Gauss, C. F. (1809). *Theoria motus corporum coelestum.* Hamburgi: Sumtibus F. Perthes et I.H. Besser.

Groves, T. and Rothenberg, T. (1969). A note on the expected value of an inverse matrix. *Biometrika*, 56(3):690–691.

Haff, L. (1979). An identity for the wishart distribution with applications. *Journal of Multivariate Analysis*, 9(4):531–544.

Han, J., Pei, J., and Kamber, M. (2011). *Data mining: concepts and techniques.* Amsterdam: Elsevier.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The elements of statistical learning, 2nd edition.* New York, NY: Springer.

Hastie, T., Tibshirani, R., and Wainwright, M. (2015). *Statistical learning with sparsity: the lasso and generalizations.* Boca Raton, FL: CRC press.

Hocking, R. and Leslie, R. (1967). Selection of the best subset in regression analysis. *Technometrics*, 9(4):531–540.

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

Kloft, M., Brefeld, U., Laskov, P., Müller, K. R., Zien, A., and Sonnenburg, S. (2009). Efficient and accurate lp-norm multiple kernel learning. In *Advances in neural information processing systems 22*, pages 997–1005.

Laplace, P. S. (1810). *Mémoire sur les approximations des formules qui sont fonctions de très-grands nombres, et sur leur application aux probabilités.* París: Baudouin.

Larose, D. T. and Larose, C. D. (2015). *Data mining and predictive analytics*. Hoboken, NJ: John Wiley & Sons.

Legendre, A. M. (1805). *Nouvelles méthodes pour la détermination des orbites des comètes*. París: F. Didot.

Leskovec, J., Rajaraman, A., and Ullman, J. D. (2014). *Mining of massive datasets*. Cambridge, UK: Cambridge University Press.

Montgomery, D. C., Peck, E. A., and Vining, G. G. (2012). *Introduction to linear regression analysis, 5th edition*. Hoboken, NJ: John Wiley & Sons.

Nisbet, R., Miner, G., and Elder IV, J. (2009). *Handbook of statistical analysis and data mining applications*. London: Academic Press.

Osborne, M. R., Presnell, B., and Turlach, B. A. (2000). A new approach to variable selection in least squares problems. *IMA journal of numerical analysis*, 20(3):389–403.

Stigler, S. M. (1981). Gauss and the invention of least squares. *The Annals of Statistics*, 9(3):465–474.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288.

Tibshirani, R. (2017). *Machine Learning 10-702: Sparsity, the Lasso, and Friends*. Retrieved from: http://www.stat.cmu.edu/ ryantibs/statml/lectures/sparsity.pdf.

Tibshirani, R. J. et al. (2013). The lasso problem and uniqueness. *Electronic Journal of Statistics*, 7:1456–1490.

Wishart, J. (1928). The generalised product moment distribution in samples from a normal multivariate population. *Biometrika*, 20A(1/2):32–52.

Wu, H. C. (2007). The Karush-Kuhn-Tucker optimality conditions in an optimization problem with interval-valued objective function. *European Journal of Operational Research*, 176(1):46–59.

Yan, X. and Su, X. (2009). *Linear regression analysis: theory and computing*. Singapore: World Scientific.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320.

Anexo I

Problemas de Optimización Convexos

Un **conjunto** $C \subseteq \mathbb{R}^n$ se denomina **convexo** si para cualquier $\boldsymbol{x}$, $\boldsymbol{y} \in C$ y $t \in [0,1]$ se verifica que

$$t\boldsymbol{x} + (1-t)\boldsymbol{y} \in C, \tag{5.1}$$

es decir, si para cualquier $\boldsymbol{x}$, $\boldsymbol{y} \in C$ el segmento que une $\boldsymbol{x}$ e $\boldsymbol{y}$ se encuentra completamente dentro de C.

Una **función** $f : \mathbb{R}^n \to \mathbb{R}$ se denomina **convexa** si su dominio $Dom(f)$ es convexo, y para cualquier $\boldsymbol{x}$, $\boldsymbol{y} \in Dom(f)$ y $t \in [0,1]$ se verifica que

$$f(t\boldsymbol{x} + (1-t)\boldsymbol{y}) \leq tf(\boldsymbol{x}) + (1-t)f(\boldsymbol{y}), \tag{5.2}$$

es decir, si para cualquier $\boldsymbol{x}$, $\boldsymbol{y} \in Dom(f)$ la función se encuentra debajo del segmento lineal que une la función evaluada en $\boldsymbol{x}$ e $\boldsymbol{y}$. Una función es estrictamente convexa si esta desigualdad se cumple estrictamente para $\boldsymbol{x} \neq \boldsymbol{y}$ y $t \in (0,1)$. La opuesta de una función convexa es una función cóncava.

Las funciones afines son las únicas funciones que son convexas y cóncavas simultáneamente (Boyd and Vandenberghe, 2004). Una función $G : \mathbb{R}^m \to \mathbb{R}^n$ es **afín** si existen una función lineal $L : \mathbb{R}^m \to \mathbb{R}^n$ y un vector $\boldsymbol{b}$ en $\mathbb{R}^n$ tales que $G(\boldsymbol{x}) = L(\boldsymbol{x}) + \boldsymbol{b}$ para todo $\boldsymbol{x}$ en $\mathbb{R}^m$. Una función $L : \mathbb{R}^m \to \mathbb{R}^n$ es **lineal** si se verifica que para cualquier par de vectores $\boldsymbol{x}$ e $\boldsymbol{y}$ en $\mathbb{R}^m$, $L(\boldsymbol{x} + \boldsymbol{y}) = L(\boldsymbol{x}) + L(\boldsymbol{y})$, y que para cualquier vector $\boldsymbol{x}$ en $\mathbb{R}^m$ y escalar a, $L(a\boldsymbol{x}) = aL(\boldsymbol{x})$. En otras palabras, una función afín es simplemente una función lineal más una traslación.

Si $G : \mathbb{R}^m \to \mathbb{R}^n$ es afín, entonces existen una matriz $\boldsymbol{M}$ de dimensión $n \times m$ y un vector $\boldsymbol{b}$ en $\mathbb{R}^n$ tales que $A(\boldsymbol{x}) = \boldsymbol{M}\boldsymbol{x} + \boldsymbol{b}$, para todo $\boldsymbol{x}$ en $\mathbb{R}^m$. En particular, si $f : \mathbb{R} \to \mathbb{R}$ es afín, entonces existen números reales m y b tales que $f(x) = mx + b$, para todo x en $\mathbb{R}$.

Formalmente, un problema de minimización como los planteados para el cálculo de esti-

madores en problemas de regresión tiene la forma

$$\begin{aligned} &\min_{\boldsymbol{x}\in D} \ f(\boldsymbol{x}) \\ &\text{sujeto a } \ h_i(\boldsymbol{x}) \leq 0, \ \ i = 1, ..., m \\ &\qquad\quad \ell_j(\boldsymbol{x}) = 0, \ \ j = 1, ..., r, \end{aligned}$$

donde D es el dominio común de todas las funciones. A f se la llama función objetivo o criterio. Un punto factible $\boldsymbol{x}$ es un punto en D que verifica todas las restricciones de igualdad y desigualdad. Una solución o minimizador $\boldsymbol{x}^*$ es un punto factible que alcanza el menor valor del criterio. Usualmente, al menor valor del criterio se lo denota con f^*.

Un problema de optimización **convexo** es aquel donde todas las funciones f, h_1, ..., h_m son convexas, y todas las funciones ℓ_1, ..., ℓ_r son afines, es decir, $\ell_j(\boldsymbol{x}) = \boldsymbol{a}_j^T\boldsymbol{x} + b_j$.

En estos problemas de optimización se requiere que las funciones ℓ_j sean afines porque la igualdad $\ell_j(\boldsymbol{x}) = 0$ puede escribirse como un conjunto de dos desigualdades simultáneas ($\ell_j(\boldsymbol{x}) \leq 0$ y $-\ell_j(\boldsymbol{x}) \leq 0$) y, como $\ell_j(\boldsymbol{x})$ y $-\ell_j(\boldsymbol{x})$ deben ser funciones convexas, ℓ_j debe ser convexa y cóncava simultáneamente, es decir, debe ser una función afín.

Un problema de optimización es **estrictamente convexo** cuando alguna de las funciones f, h_1, ..., h_m es estrictamente convexa.

En los problemas de optimización convexos, un punto factible $\boldsymbol{x}$ es un minimizador local si para todo punto factible $\boldsymbol{y}$ tal que $||\boldsymbol{x} - \boldsymbol{y}||_2 \leq R$, con $R > 0$, se verifica que $f(\boldsymbol{x}) \leq f(\boldsymbol{y})$.

Demostramos por el absurdo que en este tipo de problemas cualquier minimizador local es un minimizador global. Sea $\boldsymbol{x}$ un minimizador local de un problema de optimización convexo. Supóngase que $\boldsymbol{x}$ no es un minimizador global, es decir, existe algún punto factible $\boldsymbol{z}$ tal que $f(\boldsymbol{z}) < f(\boldsymbol{x})$. La convexidad del dominio D y de las restricciones del problema de optimización implican que el punto $t\boldsymbol{z} + (1-t)\boldsymbol{x}$ es factible para cualquier $0 \leq t \leq 1$. Por

ser f convexa,

$$f(t\boldsymbol{z} + (1-t)\boldsymbol{x}) \leq tf(\boldsymbol{z}) + (1-t)f(\boldsymbol{x}) < f(\boldsymbol{x}) \tag{5.3}$$

para cualquier $0 \leq t \leq 1$. Entonces, se puede elegir $t > 0$ lo suficientemente pequeño de modo que $||\boldsymbol{x} - (t\boldsymbol{z} + (1-t)\boldsymbol{x})||_2 = t||\boldsymbol{x} - \boldsymbol{z}||_2 \leq R$, lo cual es absurdo, ya que $\boldsymbol{x}$ es un minimizador local.

Anexo II

Ajuste LASSO usando el paquete glmnet de R

El paquete **glmnet** ajusta modelos lineales generalizados a través de máxima verosimilitud penalizada. El algoritmo es extremadamente rápido, y puede trabajar con matrices de predictores esparcidas. Puede ajustar modelos de regresión lineal, logísticos, multinomiales, Poisson y de Cox. Por defecto, el paquete ajusta el modelo lineal (`family=gaussian` en la función `glmnet`), siendo el que es utilizado en esta tesina.

Al ajustar modelos lineales, **glmnet** resuelve el siguiente problema de minimización

$$\min_{(\beta_0,\boldsymbol{\beta})\in\mathbb{R}^{p+1}} \frac{1}{2n}\sum_{i=1}^{n}(y_i-\beta_0-x_i^T\boldsymbol{\beta})^2+\lambda^*\left[(1-\alpha)||\boldsymbol{\beta}||_2^2/2+\alpha||\boldsymbol{\beta}||_1\right],$$

donde $\lambda^* \geq 0$ es el parámetro de suavizado y $0 \leq \alpha \leq 1$ es la penalidad de la red elástica. Un valor $\alpha = 0$ equivale a ajustar una regresión *Ridge*, mientras que con $\alpha = 1$ se ajusta una regresión LASSO. En este trabajo se utilizó únicamente $\alpha = 1$, ya que el camino de soluciones *Ridge* es calculado a través de la programación de (3.37), aprovechando el hecho que la solución de este método de regularización se puede expresar en forma cerrada. Para LASSO, la relación que existe entre λ^* y λ en (3.53) es $\lambda^* = \dfrac{\lambda}{n}$.

Las soluciones se calculan para distintos valores predeterminados de λ^* que cubren todo su rango de variación $[0, \infty)$.

El algoritmo que se utiliza para calcular la solución en **glmnet** es *coordinate descent*, el cual optimiza sucesivamente la función objetivo sobre cada parámetro manteniendo a los demás fijos hasta lograr convergencia. Este algortimo calcula el camino de soluciones rápidamente al usar técnicas como *warm starts*.

El paquete incluye métodos para realizar predicciones y gráficos, así como una función para realizar validación cruzada.

Para instalar el paquete, puede escribirse el siguiente comando en la consola de R:

`install.packages(‘‘glmnet’’, repos = ‘‘http://cran.us.r-project.org’’)`.

Para utilizar el paquete, es necesario cargar la librería con el comando:
`library(glmnet)`.

Para ajustar un modelo, se utiliza la función `glmnet`:
`fit = glmnet(x, y)`.

`fit` es un objeto de clase glmnet que contiene toda la información relevante del modelo ajustado, `x` es la matriz de predictores e `y` es el vector de respuestas. Previo a ajustar el modelo, la matriz de predictores es estandarizada por defecto. Los coeficientes siempre se devuelven en la escala original. Diferentes funciones pueden usarse para ejecutar diferentes tareas, entre ellas:

- `plot(fit)`: presenta un gráfico con el camino de soluciones.
- `print(fit)`: presenta un resumen con la cantidad de coeficientes distintos de cero, el porcentaje de la deviance explicada y el valor de λ^* en los distintos pasos del camino de soluciones.
- `coef(fit, s=`λ_0`)`: muestra los coeficientes de la solución correspondiente a $\lambda^* = \lambda_0$.
- `predict(fit, newx=nx, s=`λ_0`)`: permite realizar predicciones sobre una nueva matriz de predictores `nx`, utilizando el ajuste correspondiente al valor de $\lambda^* = \lambda_0$.

El valor de λ_0 puede ser elegido por validación cruzada. Si bien no se utiliza en esta tesina, la función `cv.glmnet` realiza el ajuste aplicando validación cruzada. Al utilizar el comando `cvfit=cv.glmnet(x, y)`, `cvfit$lambda.min` es el valor de λ^* que produce el menor error al aplicar validación cruzada. Otro valor almacenado es `lambda.1se`, y es el λ^* que produce el modelo más regularizado que se encuentra dentro de un error estándar del mínimo. Para utilizar esos valores, sólo es necesario escribir `s=lambda.min` o `s=lambda.1se` en las funciones `coef` o `predict`.

Otras opciones que provee **glmnet** son:

- `nlambda`: el número de valores de λ^* en la secuencia. El valor por defecto es 100. En esta tesina, se calculó el camino LASSO para `nlambda=50`.

- `lambda`: permite calcular la solución para un valor determinado, aunque por lo general no se utiliza y el programa calcula las soluciones para una secuencia de valores. En esta tesina esta opción se usó para calcular los estimadores MC, fijando `lambda=0`.

Anexo III

Programas disponibles para la obtención de los estimadores de los métodos estudiados

En este Anexo se presentan diferentes paquetes y funciones de R para ajustar las regresiones *Ridge* y LASSO, así como las instrucciones necesarias para obtener los mismos resultados con cada uno de ellos. También se incluyen comentarios del ajuste mínimo-cuadrático utilizando el paquete **glmnet**.

El conjunto de datos (data set) utilizado para realizar los ajustes contiene la siguiente información sobre 128 deportistas de diferentes divisiones de un importante club de fútbol de Argentina:

- Identificación de cada jugador (`jugador`).
- División a la que pertenece el jugador (`div`).
- Peso del jugador expresado en kg (`peso`).
- "Diámetro biacromial", mide la distancia en cm de un lado al otro, entre los procesos acromiales derecho e izquierdo de la escápula. Provee una indicación del diámetro de los hombros (`biacro`).
- "Diámetro biileocrestídeo", mide la distancia en cm de un lado al otro, entre las partes más laterales de las crestas ilíacas. Provee una indicación del ancho de la cadera (`biilio`).
- "Diámetro transverso del tórax", mide la distancia expresada en cm de los puntos más laterales del tórax, a nivel de la cuarta costilla (`toraxt`).
- "Perímetro de brazo relajado", mide la distancia perimetral en cm del brazo derecho en ángulo recto al eje longitudinal del húmero (`brazorel`).
- "Perímetro de muslo superior", es el perímetro expresado en cm del muslo derecho (`muslosup`).

- "Perímetro de muslo medial", es la medición en cm del perímetro del muslo derecho tomada perpendicular al eje longitudinal del muslo (`muslomed`).

- "Perímetro de pantorrilla", es el máximo perímetro de la pantorrilla. En este caso la medición ha sido dicotomizada. Si el diámetro es menor a 36 cm, toma el valor 0, caso contrario toma el valor 1 (`panto`).

La variable respuesta es el peso del jugador. Las mediciones antropométricas fueron realizadas en el año 2012 durante una consulta con un nutricionista. Se trabaja únicamente con las variables continuas, es decir, se excluyen del análisis tanto la variable que indica la división a la que pertenece el jugador como la variable dicotomizada.

Lectura del data set.

```
DATA = read.csv(‘‘Ruta/Datos Jugadores.csv’’)
```

Creación de la matriz de predictores $\boldsymbol{X}$, sólo con las variables cuantitativas continuas.

```
X <- model.matrix(peso ~ biacro + biilio + toraxt + brazorel + muslosup +
muslomed, DATA)
```

Observación: la primer columna de $\boldsymbol{X}$ tiene todos 1.

Estandarización de la matriz $\boldsymbol{X}$.

Se elimina la columna de 1.

```
X_AUX <- X[,-1]
```

Se estandariza la matriz.

```
n <- nrow(DATA)

XEST <- scale(X_AUX) * sqrt(n/(n-1))
```

Observación: `scale` utiliza el desvío dividiendo por $n-1$. La función `lm.ridge` que se utiliza más adelante estandariza usando el desvío dividiendo por n. Por este motivo, para obtener los mismos estimadores *Ridge* a través del desarrollo teórico (multiplicación de

matrices) y usando la función `lm.ridge`, se multiplica el resultado obtenido en `scale()` por `sqrt(n/(n-1))`.

Definición del vector de respuestas.

```
Y = as.matrix(DATA[,3])
```

Cálculo del estimador *Ridge*

• Cálculo teórico a través de la multiplicación de matrices en (3.36), usando la matriz $\boldsymbol{X}$ estandarizada.

Se fija un λ para comparar los resultados de los distintos paquetes y funciones.

```
lambda = 2
```

Se calcula el vector de parámetros estimados.

```
BRTeo=solve(t(XEST)%*% XEST + lambda * diag(ncol(XEST)))%*% t(XEST)%*% Y
```

```
BRTeo
              [,1]
biacro   1.0070282
biilio   1.1268322
toraxt   2.5149678
brazorel 0.8454394
muslosup 2.7207090
muslomed 1.6791891
```

Cuando se estandarizan los predictores de la matriz $\boldsymbol{X}$, $\hat{\beta}_0$ es la media de la variable respuesta. Observación: β_0 no se penaliza. De este modo:

```
B0 = mean(DATA$peso)
```

Se calcula el vector de respuestas estimadas.

```
YestRidgeTeo = B0 + XEST%*% BRTeo
```

• Cálculo con la función `lm.ridge`, usando la matriz $\boldsymbol{X}$ sin estandarizar.

Se carga la librería MASS.

```
library(MASS)
```

Se calcula el estimador usando la función `lm.ridge`.

```
lm.ridgeDATA = lm.ridge(peso ~ biacro + biilio + toraxt + brazorel +
muslosup + muslomed, DATA, lambda = 2)
```

Se almacena el estimador y se calculan los valores predichos.

```
Blm.ridgeDATA = lm.ridgeDATA$coef
```

```
Blm.ridgeDATA
   biacro    biilio    toraxt  brazorel  muslosup  muslomed
1.0070282 1.1268322 2.5149678 0.8454394 2.7207090 1.6791891
```

```
Yestlm.ridgeDATA = scale(DATA[,c(4:9)], center = lm.ridgeDATA$xm,
scale = lm.ridgeDATA$scales) %*% lm.ridgeDATA$coef + lm.ridgeDATA$ym
```

Para estimar la respuesta con nuevas observaciones, se usan los valores `center` y `scale` del conjunto de datos de entrenamiento. La función que hay que utilizar para una matriz de predictores de datos de test previamente definida (`data.test`) es:

```
y.pred.ridge = scale(data.test, center = lm.ridgeDATA$xm,
scale = lm.ridgeDATA$scales) %*% lm.ridgeDATA$coef + lm.ridgeDATA$ym
```

• Cálculo de los valores predichos usando la DVS de la matriz $\boldsymbol{X}$ estandarizada.

```
DVS = svd(XEST)
```

```
YestDVSRidge=B0 + DVS$u%*% diag(DVS$d)%*% solve(diag(DVS$d)%*% diag(DVS$d) +
lambda * diag(nrow(diag(DVS$d)))) %*% diag(DVS$d) %*% t(DVS$u) %*% Y
```

Comparación de los valores predichos con estos tres enfoques.

```
Yest = cbind(YestRidgeTeo, Yestlm.ridgeDATA, YestDVSRidge)

head(Yest)
```

```
      [,1]     [,2]     [,3]
1 65.16657 65.16657 65.16657
2 66.18151 66.18151 66.18151
3 66.61484 66.61484 66.61484
4 50.59120 50.59120 50.59120
5 72.14322 72.14322 72.14322
6 53.14305 53.14305 53.14305
```

Observación: con los tres enfoques se obtienen los mismos valores predichos.

• Cálculo usando la programación de (3.37) que se presenta en Bien (2016), sobre la matriz $\boldsymbol{X}$ estandarizada.

Definición de la función para obtener el estimador *Ridge*.

```
ridge <- function(XEST, Y, lambda = NULL) {
                        sv <- svd(XEST)
                        df_fun <- function(lam) {
                          # degrees of freedom when tuning param is lam
                          sum(sv$d^2 / (sv$d^2 + lam))
                        }
                         df <- sapply(lambda, df_fun)
                        beta <- sapply(lambda, function(r) {
                          d <- sv$d / (sv$d^2 + r)
                          return(sv$v %*% (d * crossprod(sv$u, Y)))
                        })
                        list(beta = beta, yhat = XEST %*% beta,
                             lambda = lambda, df = df)
}
```

Se obtiene el estimador.

```
RidgeSimR <- ridge(XEST = XEST, Y = Y, lambda = 2)
```

```
RidgeSimR$beta
          [,1]
[1,] 1.0070282
[2,] 1.1268322
[3,] 2.5149678
[4,] 0.8454394
[5,] 2.7207090
[6,] 1.6791891
```

Observación: el vector de parámetros estimados obtenido es idéntico al de los procedimientos anteriores. Esta función también calcula los grados de libertad efectivos del ajuste, y es posible acceder a esta cantidad a través de `RidgeSimR$df`.

◇ Análisis de las funciones `df_fun` y `get_lam`.

Primero se calcula la DVS de la matriz de predictores estandarizada, y luego se recuerda cómo están programadas estas funciones.

```
sv <- svd(XEST)

df_fun <- function(lam) {
  sum(sv$d^2 / (sv$d^2 + lam))
}

get_lam <- function(target_df) {
  f <- function(lam) df_fun(lam) - target_df
  uniroot(f, c(0, 100 * max(sv$d^2)))$root
}
```

Se calcula la función `get_lam` para un valor igual a 5,659621.

```
get_lam(5.659621)
```

Observación: la función `get_lam` devuelve el valor de λ para el cual se tienen los grados de libertad (DF) indicados en la función. En este caso, el resultado es 2, lo que tiene sentido ya que 5,659621 son los DF con $\lambda = 2$.

El resultado de `get_lam(6)` es 0, lo cual tiene sentido ya que con 6 DF no hay penalización (en este ejemplo en el que hay 6 variables explicativas).

◇ En `lambda <- sapply(seq(1,nrow(model$x),length=nlambda), get_lam)` los DF (primer término de la función `sapply`) llegan hasta el número de filas de la matriz $\boldsymbol{X}$ porque en las simulaciones se usan matrices esparcidas ($n << p$). Esta función hace que se calcule el camino de soluciones para todo el rango de variación de λ.

◇ Comentarios sobre la función usada para el cálculo del estimador de β.

```
#Beta Ridge usando DVS
beta=solve(sv$v %*%diag(sv$d)^2 %*% t(sv$v) + lambda *diag(ncol(XEST)))
%*% sv$v %*% diag(sv$d) %*% t(sv$u) %*% Y

#En la función que se presenta en Bien (2016)
beta = sv$v %*% ((sv$d / (sv$d^2 + lambda)) * crossprod(sv$u, Y))

#Observación: en R, multiplicar (sv$d / (sv$d^2 + lambda)) -Vector-
usando "*" multiplica "1 a 1", lo cual equivale a multiplicar
diag(sv$d / (sv$d^2 + lambda)) -Matriz Diagonal- usando "%*%"

Es decir:
     sv=svd(XEST)
     d = sv$d / sv$d^2
          d * t(sv$u)%*%Y     =     diag(d) %*% t(sv$u)%*%Y
```

• Cálculo usando el paquete **glmnet**.

Cuando se calcula el estimador *Ridge* usando el paquete **glmnet**, hay que tener en cuenta la relación que existe entre el λ teórico utilizado hasta el momento y el λ del paquete, para lo cual es necesario el cálculo del desvío de la variable respuesta dividiendo por n:

```
# lambda_glmnet = sd_y * lambda / N
```

```
# Calculo Desvío de Y

sd_y <- sqrt(var(Y)*(n-1)/n)[1,1]
```

◇ Cálculo del estimador usando la matriz de predictores sin estandarizar. Comparación con el estimador teórico utilizando la misma matriz.

```
#Beta Ridge - Matrices Sin Estandarizar -


#Teórico
BR_NoEstTeo <- solve(t(X_AUX)%*%X_AUX+lambda*diag(ncol(X_AUX)))%*%
t(X_AUX)%*%(Y)

BR_NoEstTeo
                [,1]
biacro    0.05211576
biilio   -0.77370896
toraxt    0.88926071
brazorel  0.90417845
muslosup  1.30429250
muslomed -0.70749276


#Paquete glmnet
glmnet_BR_NoEst <- glmnet(X_AUX, Y, alpha = 0, standardize = F,
intercept = F, thresh = 1e-30, lambda = sd_y * lambda / n)

glmnet_BR_NoEst$beta
6 x 1 sparse Matrix of class "dgCMatrix"
                  s0
biacro    0.05211576
biilio   -0.77370896
toraxt    0.88926071
brazorel  0.90417845
muslosup  1.30429250
muslomed -0.70749276
```

Observación: se obtiene el mismo estimador.

◇ Cálculo del estimador usando la matriz de predictores estandarizada. Comparación con el estimador teórico utilizando la misma matriz.

```
#Beta Ridge - Matriz X Estandarizada -


#Teórico
BR_XEstTeo = solve(t(XEST) %*% XEST + lambda * diag(ncol(XEST))) %*%
t(XEST) %*% Y

BR_XEstTeo
              [,1]
biacro   1.0070282
biilio   1.1268322
toraxt   2.5149678
brazorel 0.8454394
muslosup 2.7207090
muslomed 1.6791891

#Paquete glmnet
glmnet_BR_XEst <- glmnet(XEST, Y, alpha = 0, standardize = F,
intercept = F, thresh = 1e-30, lambda = sd_y * lambda / n)

glmnet_BR_XEst$beta
6 x 1 sparse Matrix of class "dgCMatrix"
                s0
biacro   1.0070282
biilio   1.1268322
toraxt   2.5149678
brazorel 0.8454394
muslosup 2.7207090
muslomed 1.6791891
```

Observación: se obtiene el mismo estimador. Resulta necesario fijar

```
x = XEST, standardize = F, intercept = F
```

es decir, usar de input la matriz estandarizada "a mano", no usar la opción `standardize`.

Las comparaciones realizadas fueron para mostrar cómo obtener los mismos resultados para las distintas matrices. Como ya se mencionó en este trabajo, se recomienda el uso de la matriz estandarizada.

Cálculo del estimador LASSO

• Cálculo usando la función `lars` sobre la matriz de predictores estandarizada.

Se carga la librería.

```
library(lars)
```

Se realiza el ajuste LASSO.

```
lasso_lars <- lars(x = XEST, y=Y, normalize = FALSE,
max.steps=1000, use.Gram=FALSE)
```

Algunas opciones que ofrece el paquete:

```
lasso_lars$beta #El último Renglón es el estimador MC!
lasso_lars$lambda
lasso_lars$df
summary(lasso_lars)
plot(lasso_lars, breaks = F)
```

• Cálculo usando la función `glmnet` sobre la matriz de predictores estandarizada.

Se carga la librería.

```
library(glmnet)
```

Se realiza el ajuste LASSO.

```
lasso_glmnet <- glmnet(x = XEST, y=Y, standardize = FALSE, thresh=1e-30))
```

Algunas opciones que ofrece el paquete:

```
lasso_glmnet$beta
lasso_glmnet$lambda
lasso_glmnet$df
plot(lasso_glmnet, "lambda")
```

Observación: ambos métodos normalizan los datos de forma diferente, por lo que esta opción no debe ser usada para obtener los mismos resultados. Estos métodos no tienen opción para que el eje x del `plot` sea el mismo.

Verificación que con $\lambda = 0$ se obtiene el estimador MC

Se calcula el estimador *Ridge* cuando $\lambda = 0$ (sólo con uno de los métodos, ya que se mostró que con todos se obtienen los mismos resultados).

```
   #Ridge con lambda = 0

library(MASS)

lm.ridgeDATA = lm.ridge(peso ~ biacro + biilio + toraxt + brazorel +
muslosup + muslomed, DATA, lambda = 0)


   #Beta Estimado (lambda = 0)

Blm.ridgeDATA = lm.ridgeDATA$coef

Blm.ridgeDATA
   biacro    biilio    toraxt  brazorel  muslosup  muslomed
1.0157179 1.1085396 2.5539849 0.7845708 2.8811019 1.5910966
```

Se calcula el estimador MC sobre la matriz de predictores estandarizada y los valores predichos.

```
#Regresión Lineal Sobre X Estandarizada
DATA2 = as.data.frame(cbind(XEST, DATA[,3]))
lmDATA2 = lm(V7 ~ biacro + biilio + toraxt + brazorel +
muslosup + muslomed, DATA2)
BlmDATA2 = lmDATA2$coefficients

BlmDATA2
(Intercept)
 67.0609375
   biacro      biilio      toraxt    brazorel    muslosup    muslomed
1.0157179   1.1085396   2.5539849   0.7845708   2.8811019   1.5910966

YestlmDATA2 = lmDATA2$fitted.values
```

Observación: el estimador obtenido es el mismo.

Por otro lado, a continuación se muestra que cuando se utiliza la matriz de predictores sin estandarizar, el vector de parámetros estimados es diferente, pero se obtienen los mismos valores predichos.

```
#Regresión Lineal Sobre X (Sin Estandarizar)

DATA3 = as.data.frame(cbind(X[,-1], DATA[,3]))

lmDATA3 = lm(V7 ~ biacro + biilio + toraxt + brazorel +

muslosup + muslomed, DATA3)

BlmDATA3 = lmDATA3$coefficients

BlmDATA3
(Intercept)
-73.0234190
   biacro      biilio      toraxt    brazorel    muslosup    muslomed
0.2916258   0.7119128   1.2405998   0.3312642   0.7224330   0.4654235

YestlmDATA3 = lmDATA3$fitted.values

head(YestlmDATA2)
       1        2        3        4        5        6
65.20412 66.17082 66.64260 50.50507 72.24517 53.10333

head(YestlmDATA3)
       1        2        3        4        5        6
65.20412 66.17082 66.64260 50.50507 72.24517 53.10333
```

Por último, con el paquete **glmnet** también se puede obtener el estimador MC fijando $\lambda = 0$ y utilizando la matriz de predictores estandarizada:

```
### Paquete glmnet con lambda = 0 sobre matriz X estandarizada -> MC!

library(glmnet)

LMglmnet = glmnet(x=XEST, y=Y, lambda=0, thresh = 1e-20)

LMglmnet$beta
6 x 1 sparse Matrix of class "dgCMatrix"
                s0
biacro   1.0157179
biilio   1.1085396
toraxt   2.5539849
brazorel 0.7845708
muslosup 2.8811019
muslomed 1.5910966
```

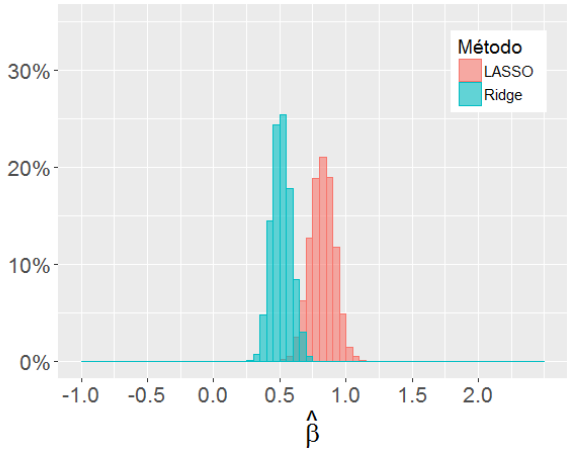
Observación: se obtiene el mismo estimador que al aplicar MC sobre la matriz de predictores estandarizada. Para una rápida verificación, se presenta nuevamente dicho estimador.

```
BlmDATA2
(Intercept)
 67.0609375
   biacro      biilio      toraxt    brazorel    muslosup    muslomed
1.0157179   1.1085396   2.5539849   0.7845708   2.8811019   1.5910966
```

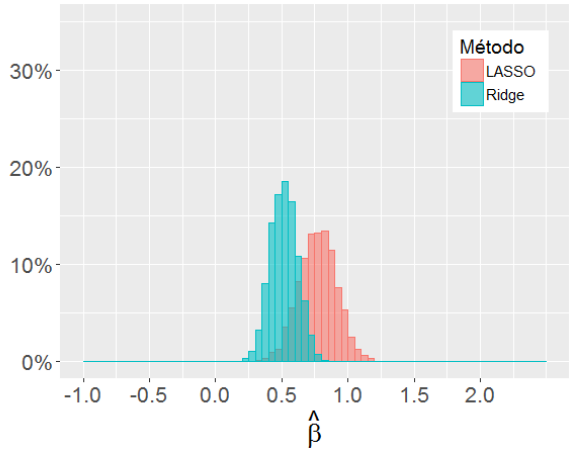
Anexo IV

Distribución Empírica de Estimadores $\hat{\beta}^{(1)}$

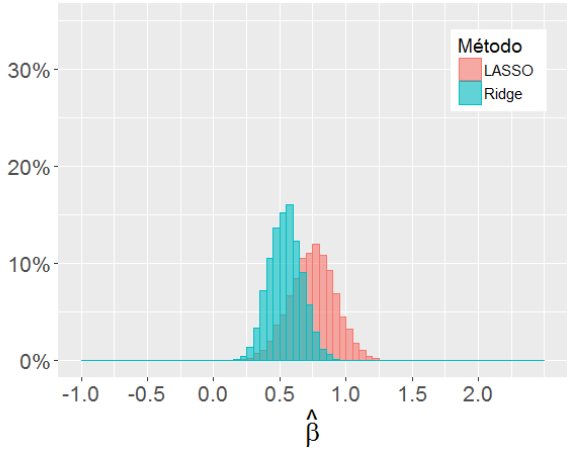
Caso $p = 100$.



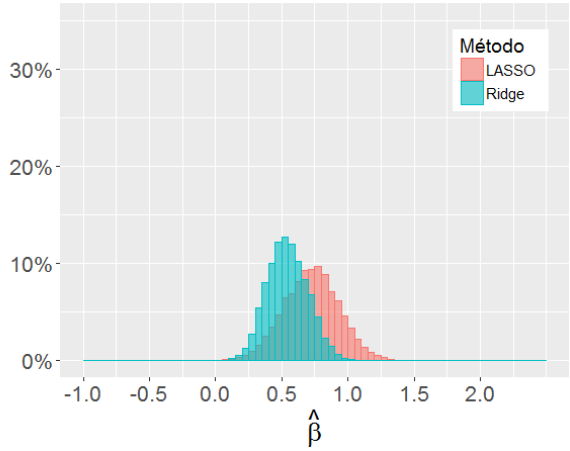
(a) $k = 2$



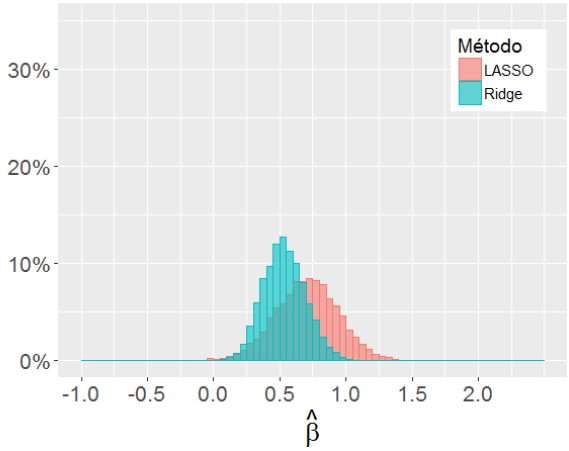
(b) $k = 4$



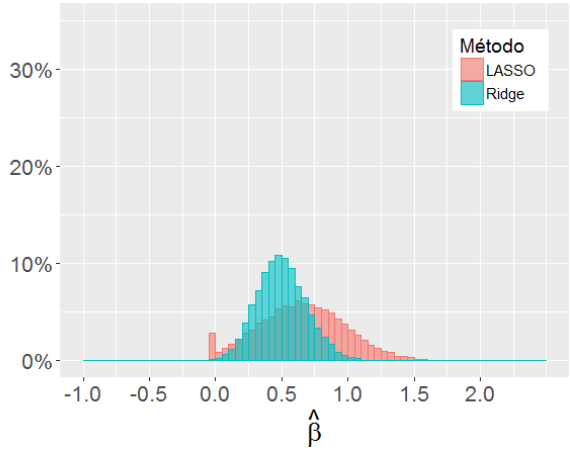
(c) $k = 6$



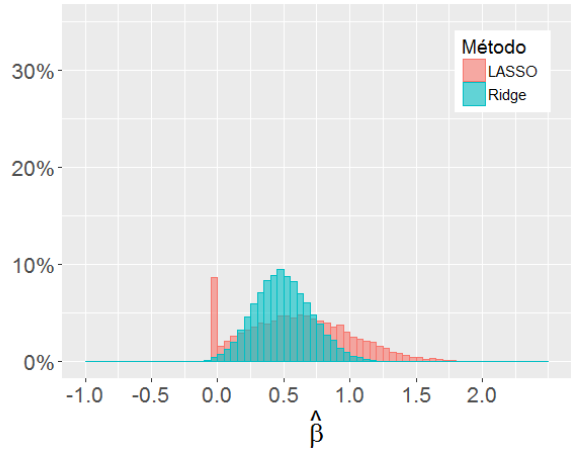
(d) $k = 8$



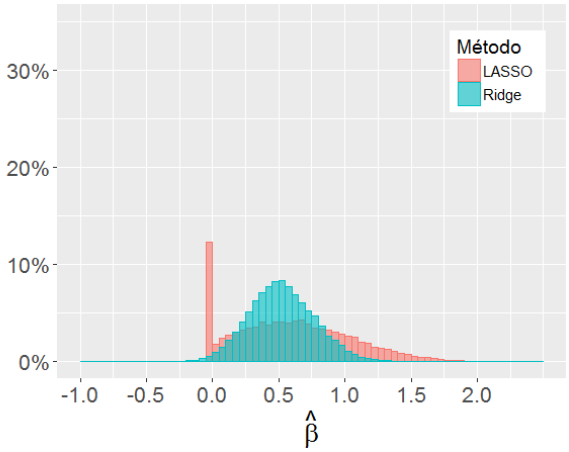
(e) $k = 10$



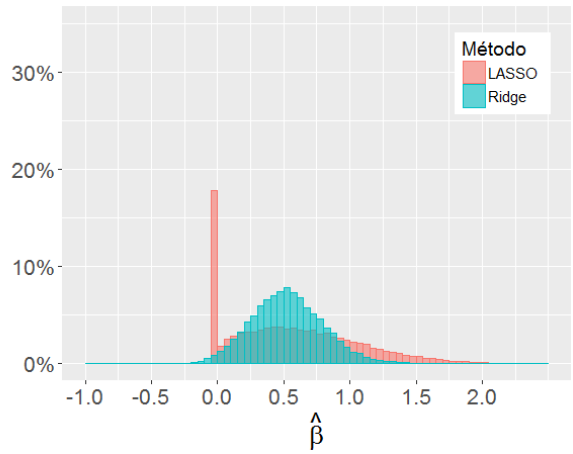
(f) $k = 15$



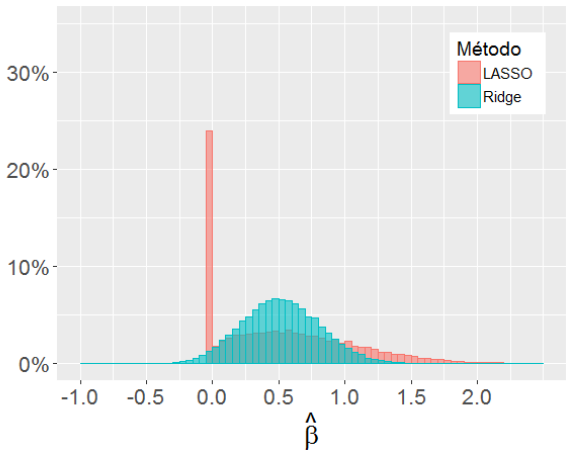
(g) $k = 20$



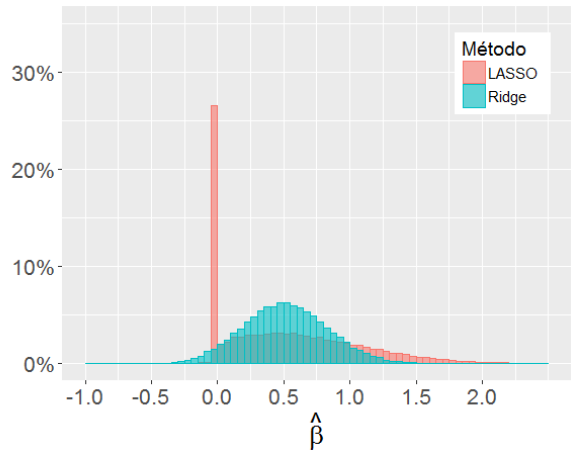
(h) $k = 25$



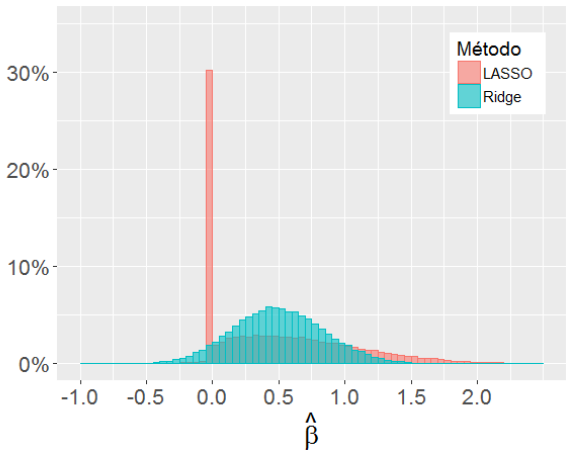
(i) $k = 30$



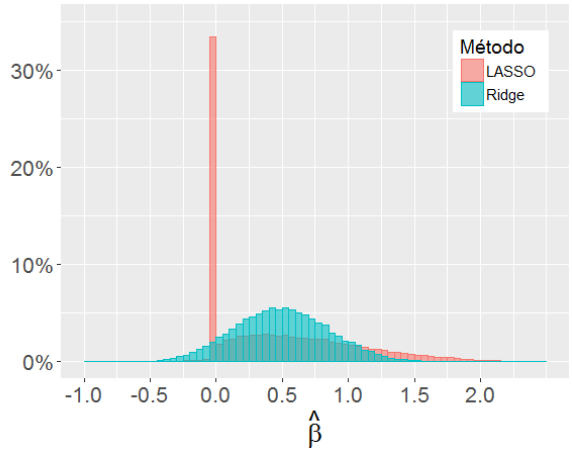
(j) $k = 35$



(k) $k = 40$

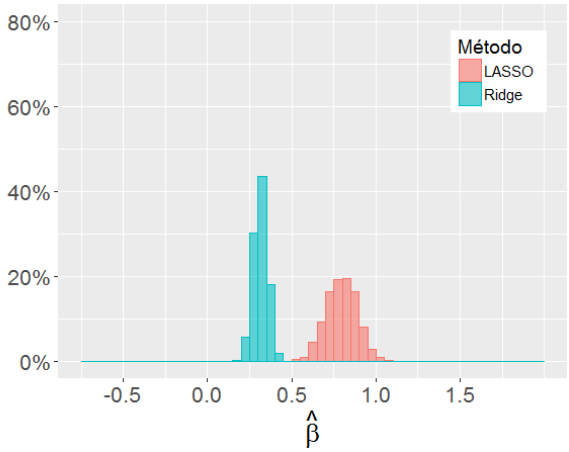


(l) $k = 45$

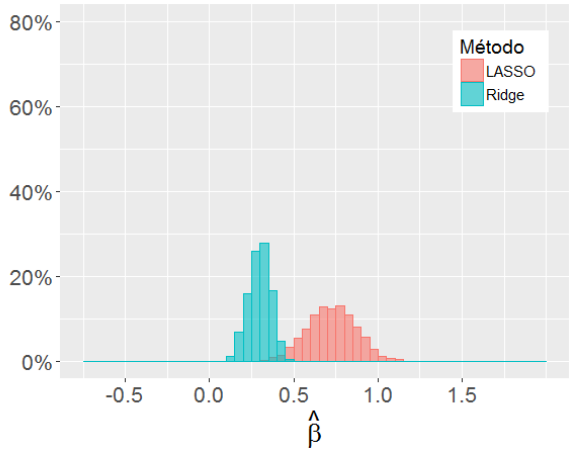


(m) $k = 50$

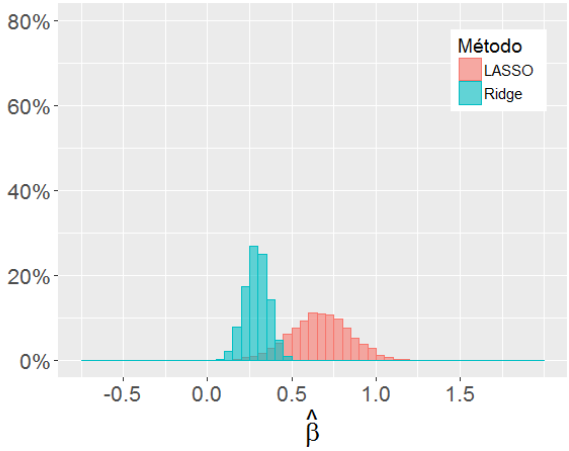
Caso $p = 200$.



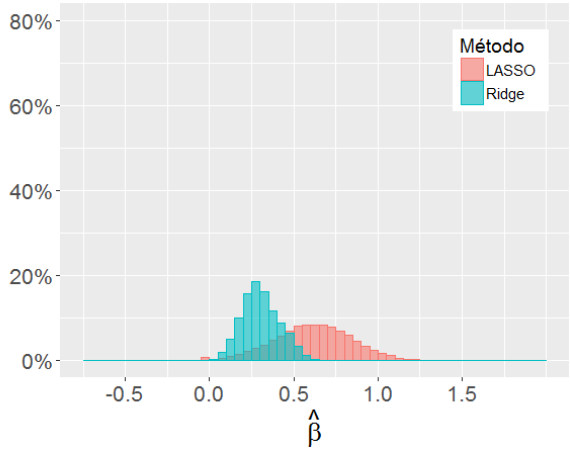
(a) $k = 2$



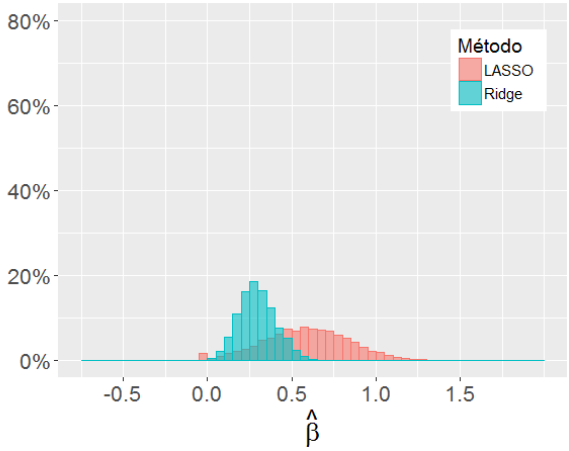
(b) $k = 4$



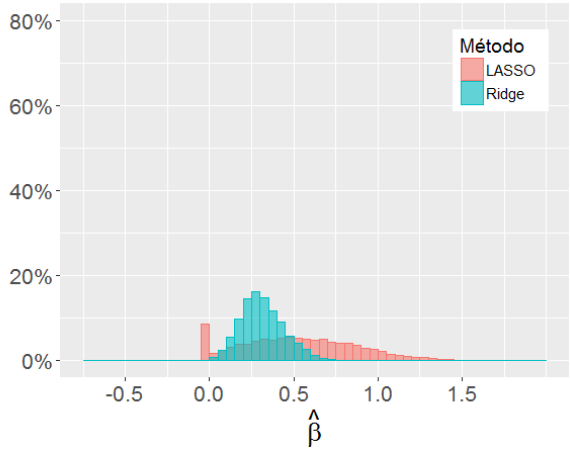
(c) $k = 6$



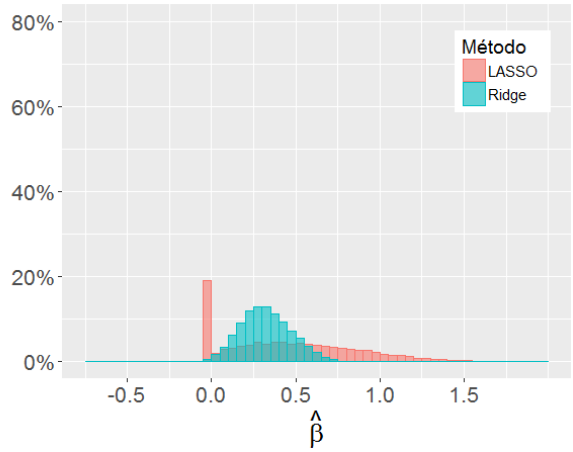
(d) $k = 8$



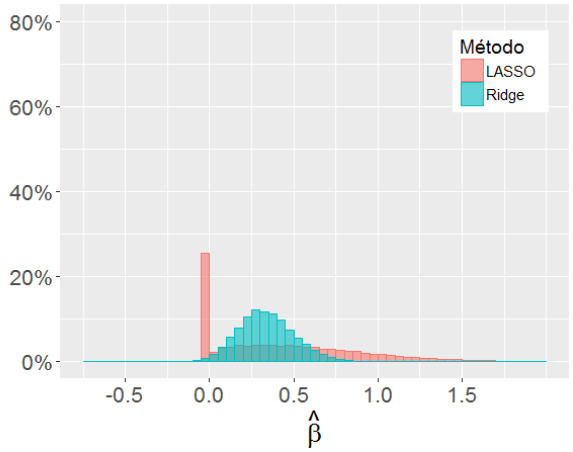
(e) $k = 10$



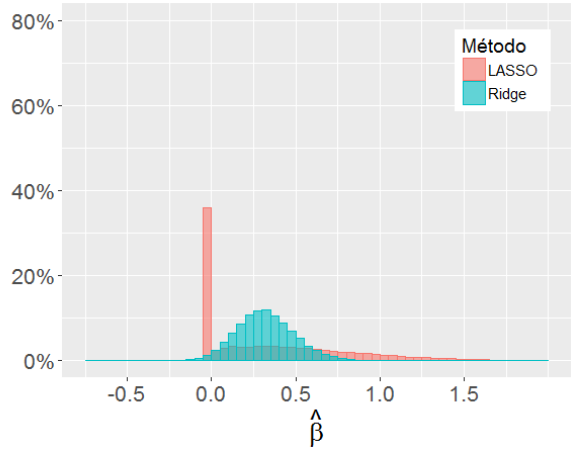
(f) $k = 15$



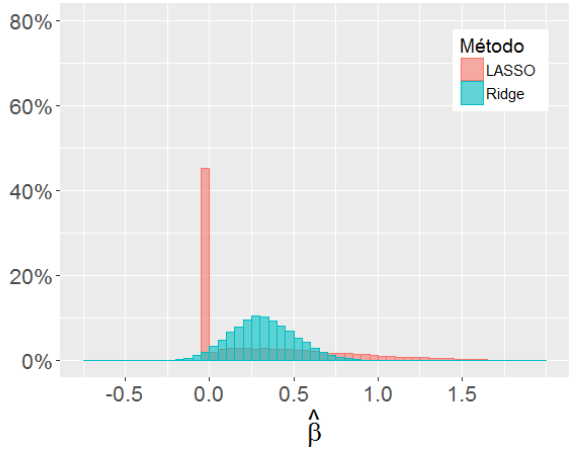
(g) $k = 20$



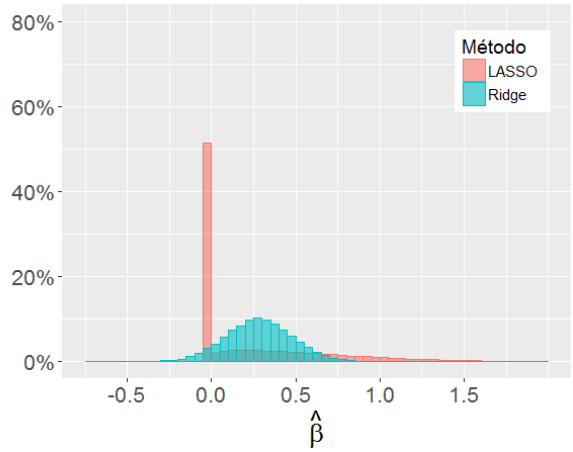
(h) $k = 25$



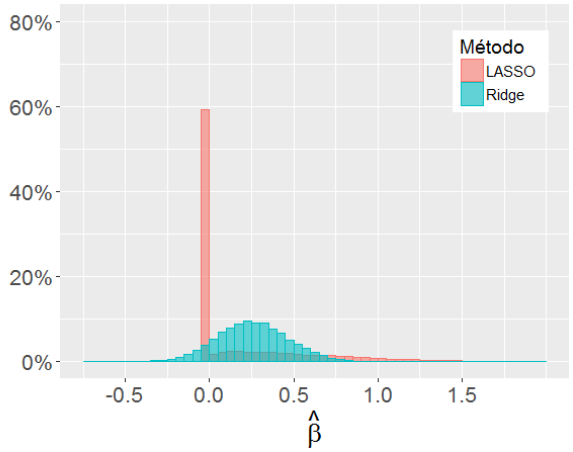
(i) $k = 30$



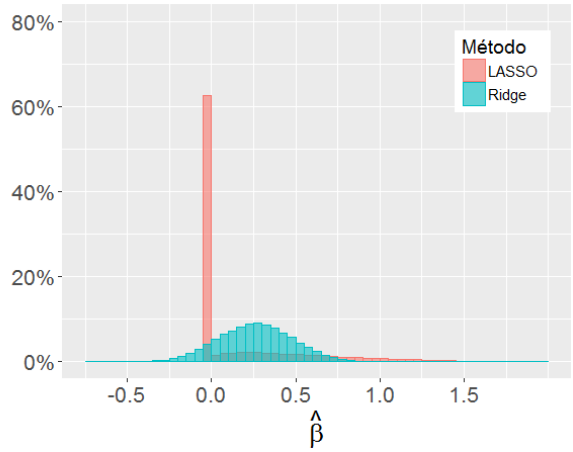
(j) $k = 35$



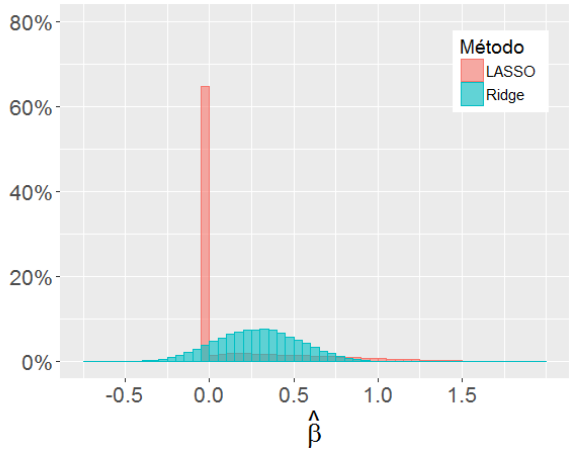
(k) $k = 40$



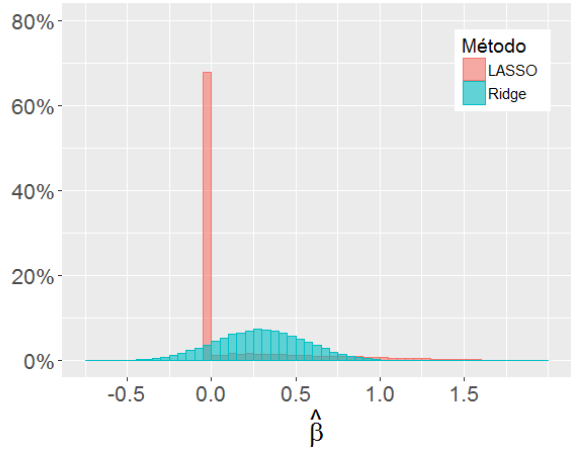
(l) $k = 45$



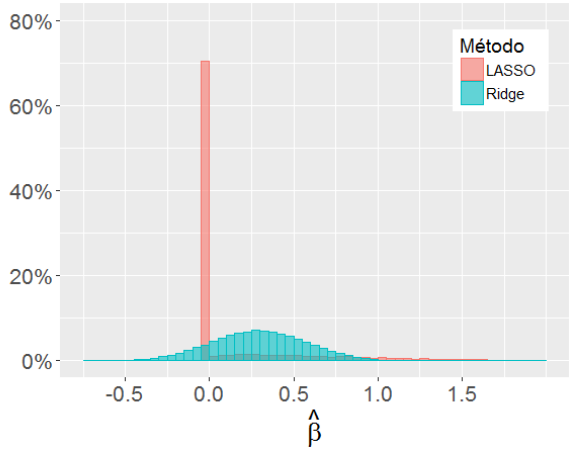
(m) $k = 50$



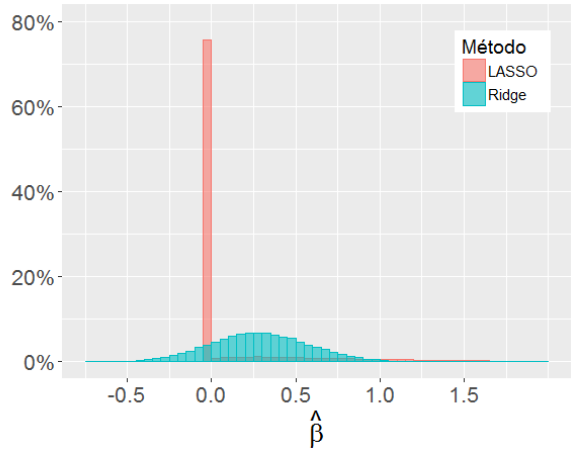
(n) $k = 60$



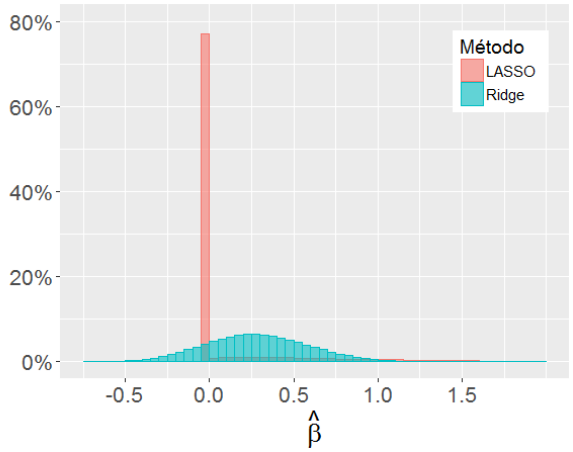
(ñ) $k = 70$



(o) $k = 80$

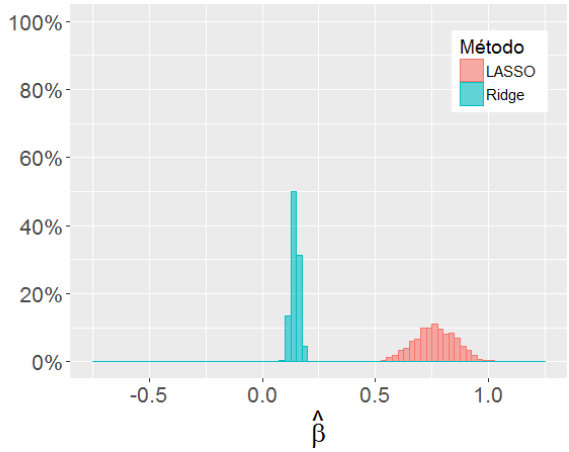


(p) $k = 90$

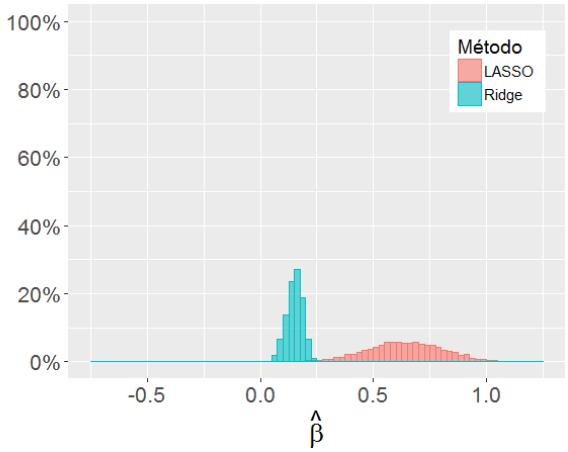


(q) $k = 100$

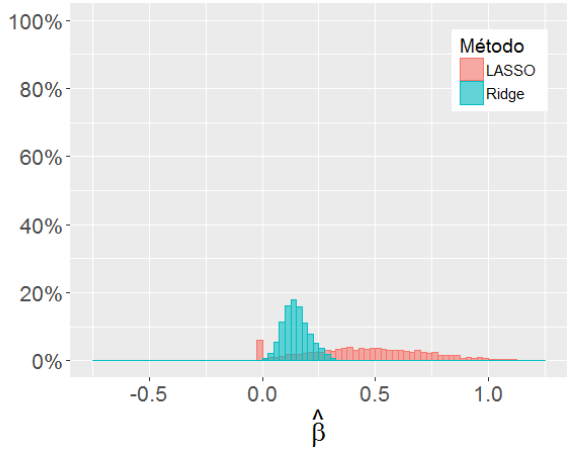
Caso $p = 400$.



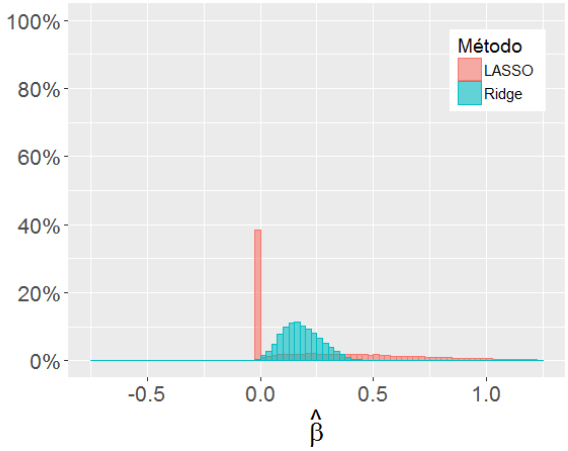
(a) $k = 2$



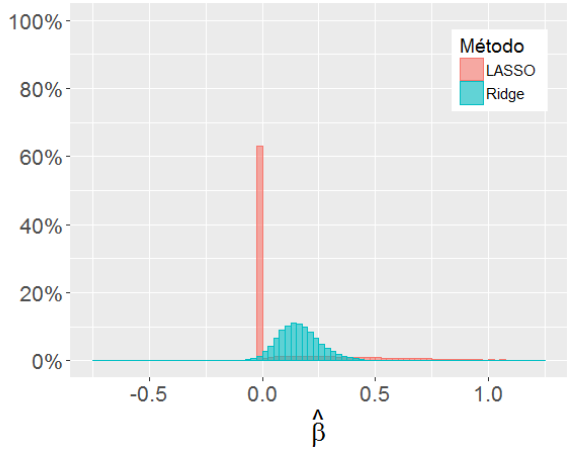
(b) $k = 5$



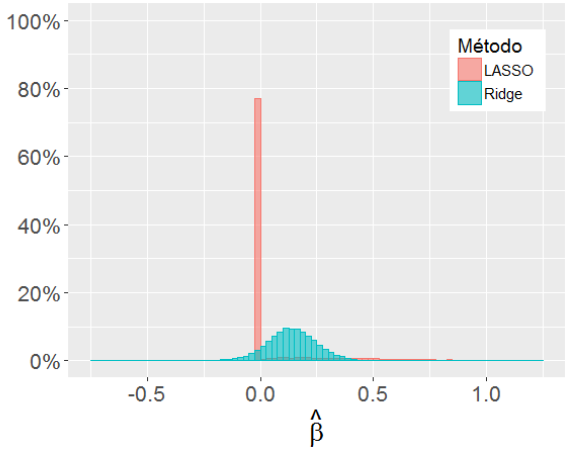
(c) $k = 10$



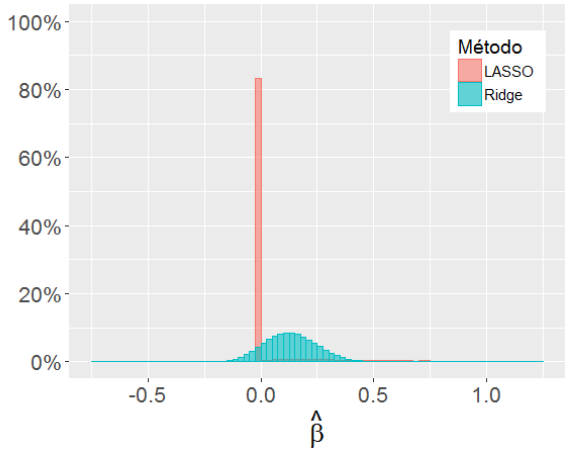
(d) $k = 20$



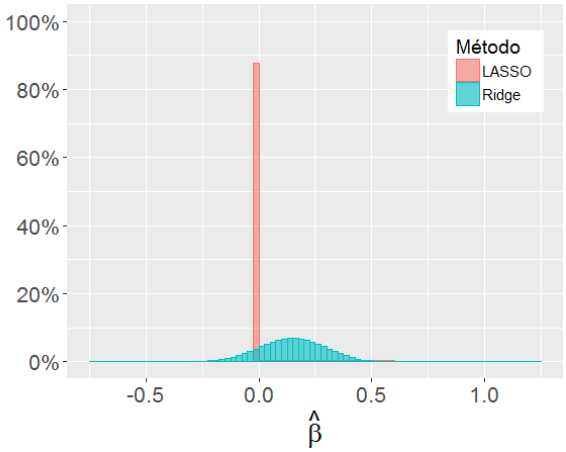
(e) $k = 30$



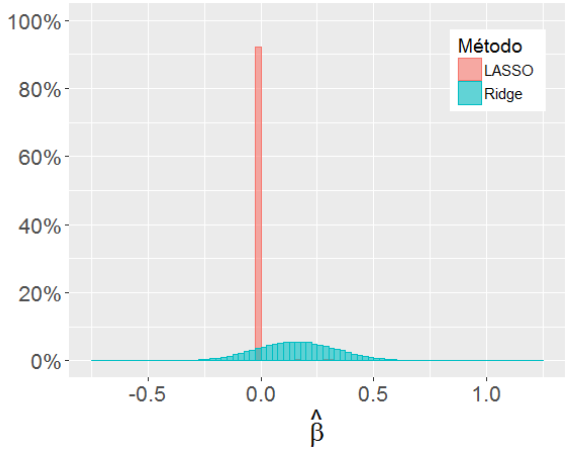
(f) $k = 40$



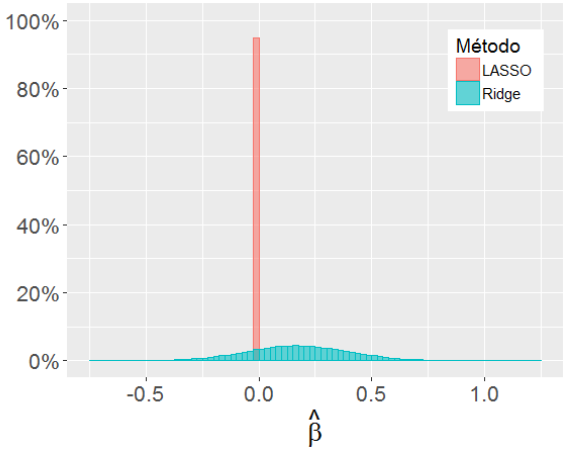
(g) $k = 50$



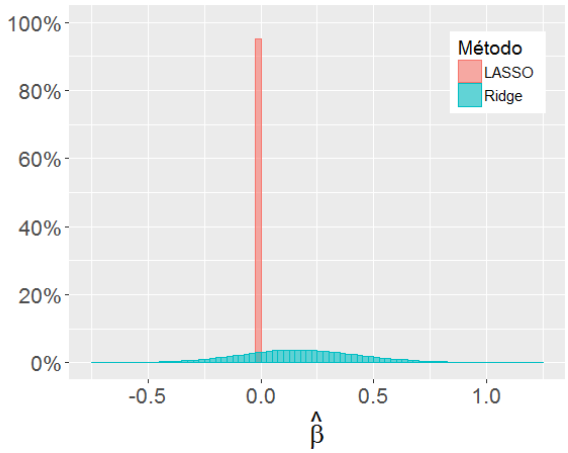
(h) $k = 75$



(i) $k = 100$



(j) $k = 150$

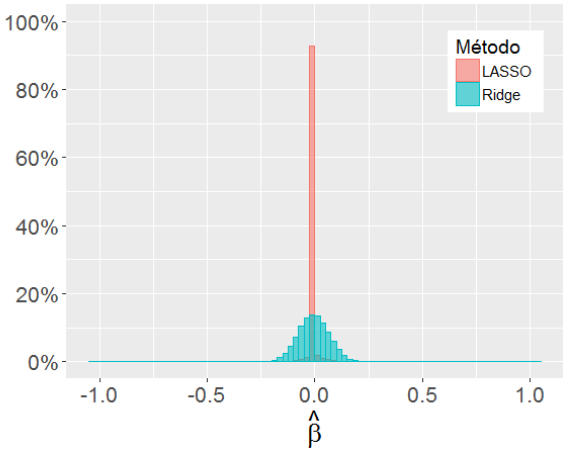


(k) $k = 200$

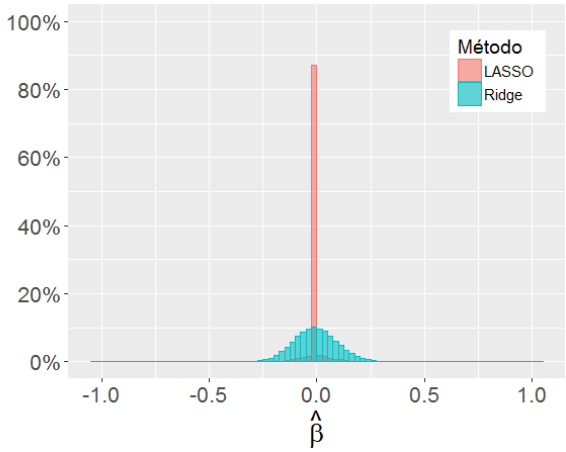
Anexo V

Distribución Empírica de Estimadores $\hat{\beta}^{(0)}$

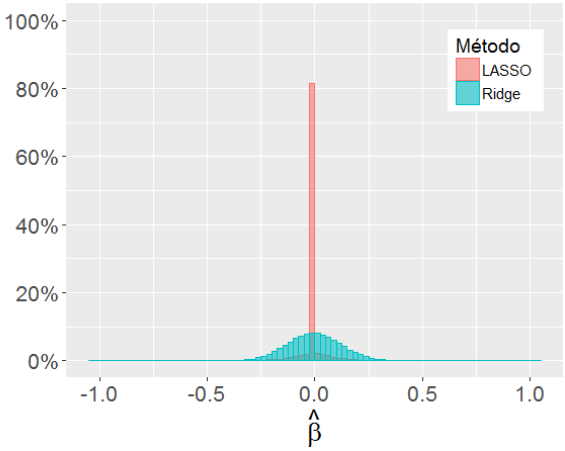
Caso $p = 100$.



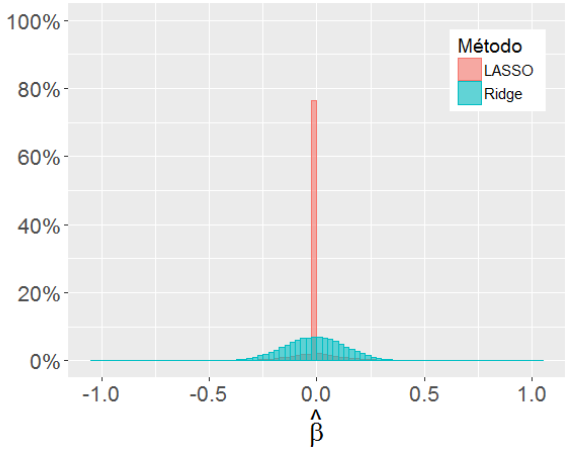
(a) $k = 2$



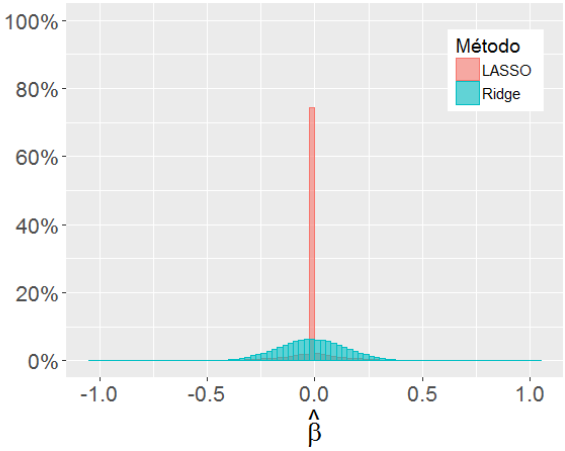
(b) $k = 4$



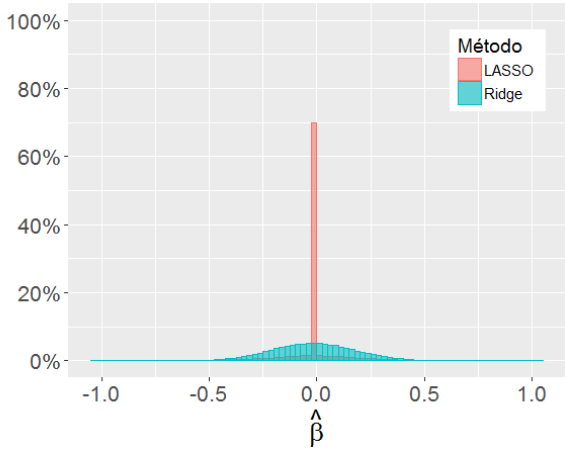
(c) $k = 6$



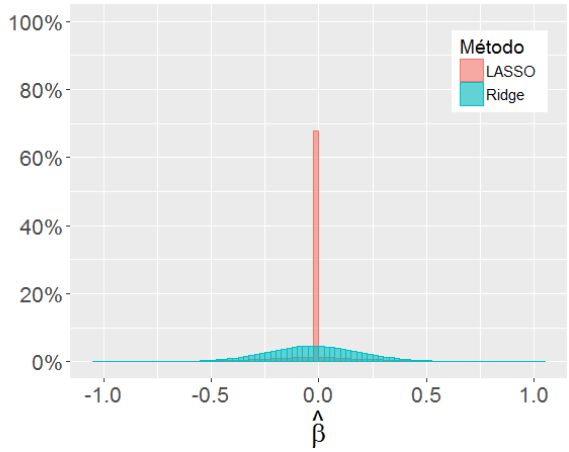
(d) $k = 8$



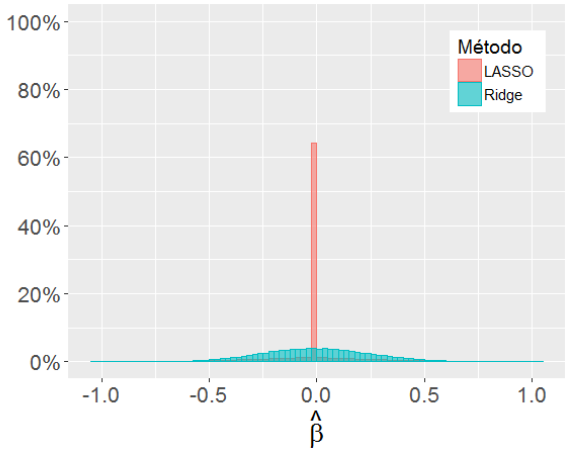
(e) $k = 10$



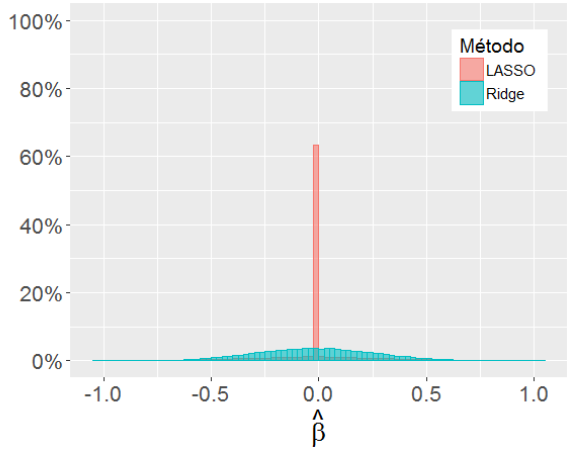
(f) $k = 15$



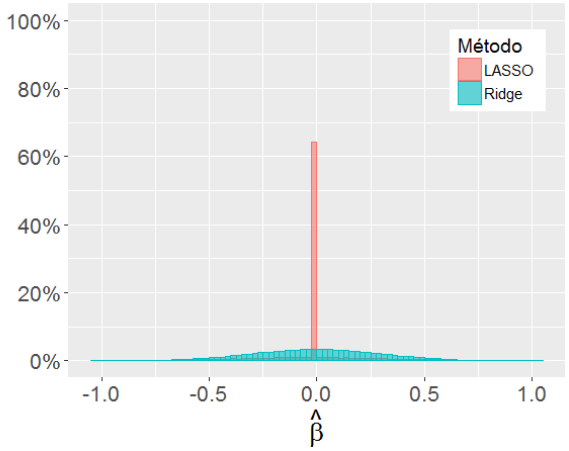
(g) $k = 20$



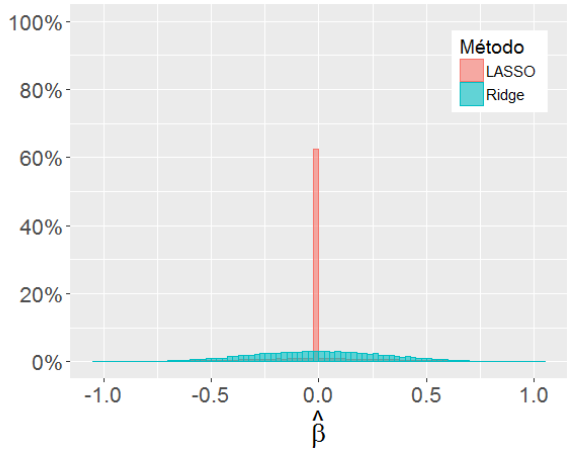
(h) $k = 25$



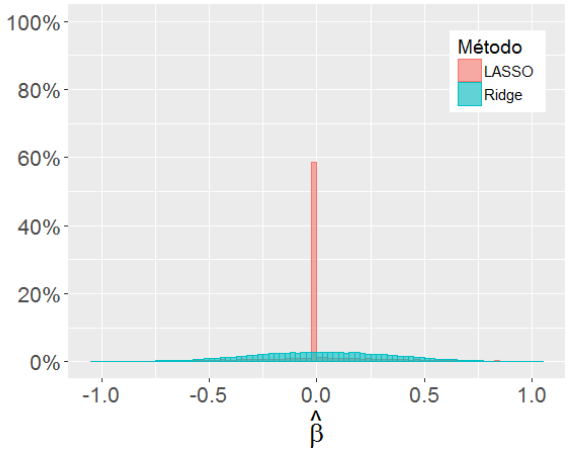
(i) $k = 30$



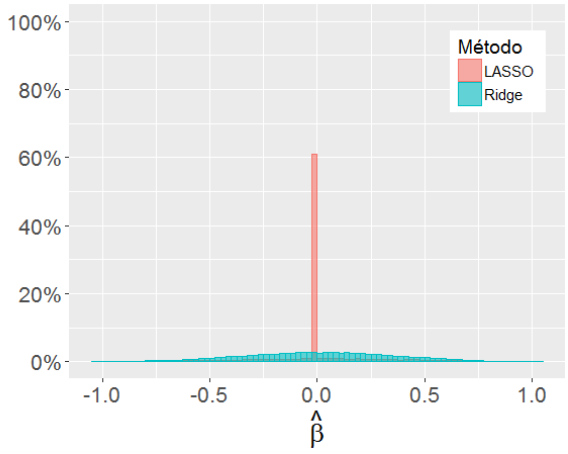
(j) $k = 35$



(k) $k = 40$

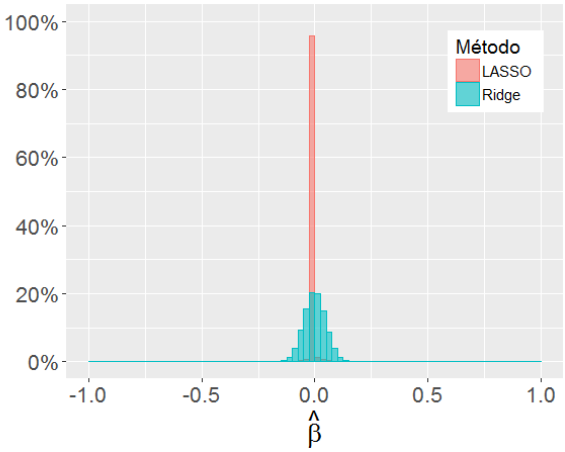


(l) $k = 45$

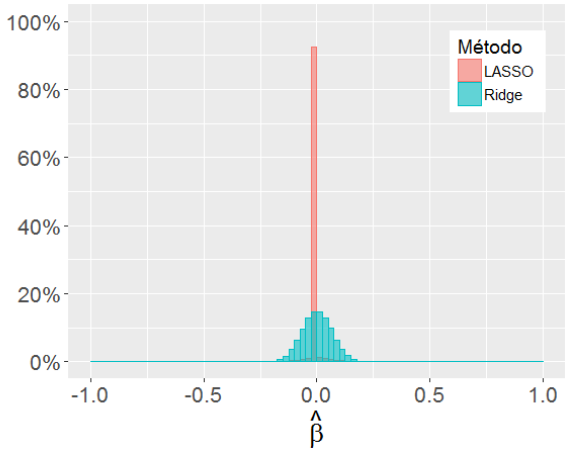


(m) $k = 50$

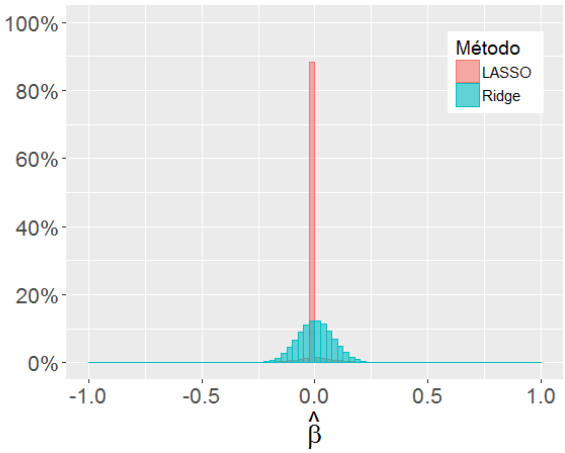
Caso $p = 200$.



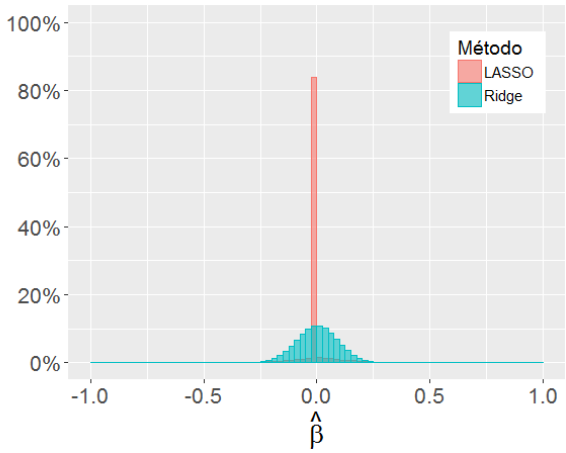
(a) $k = 2$



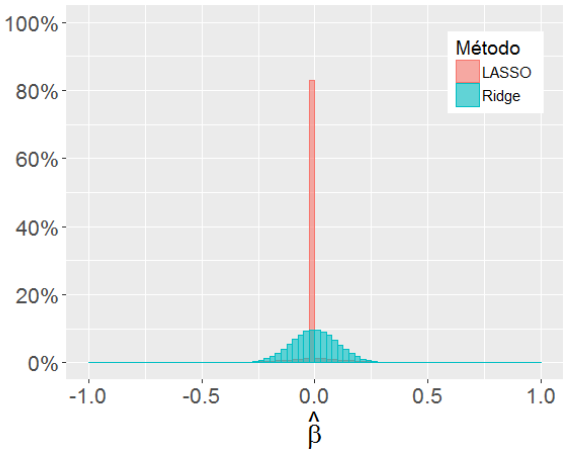
(b) $k = 4$



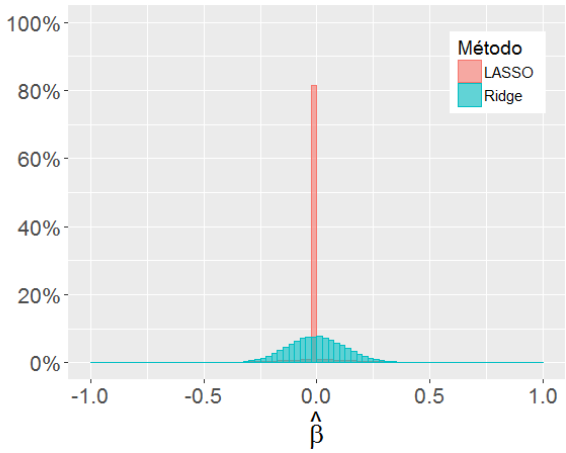
(c) $k = 6$



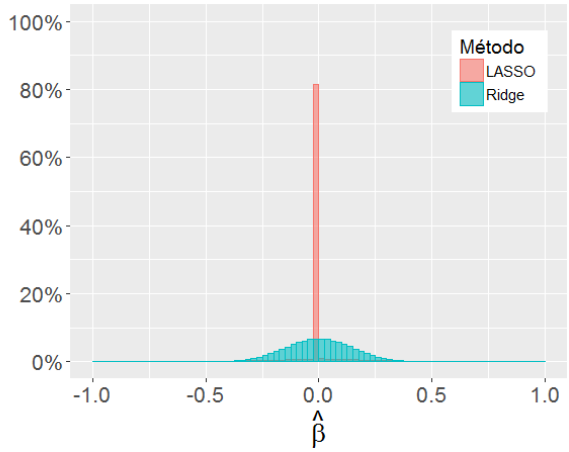
(d) $k = 8$



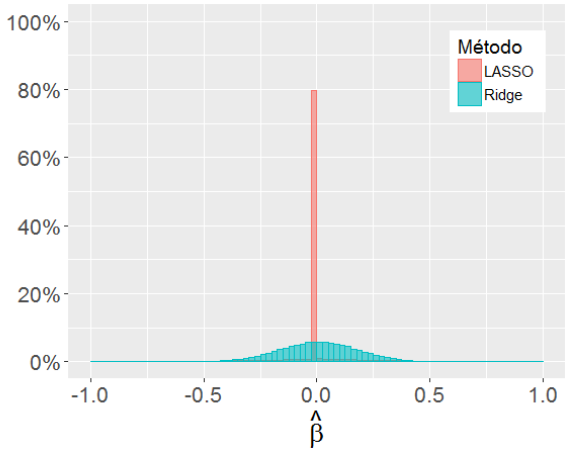
(e) $k = 10$



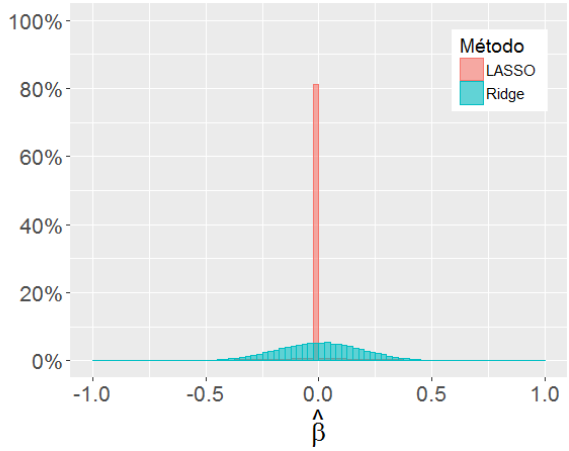
(f) $k = 15$



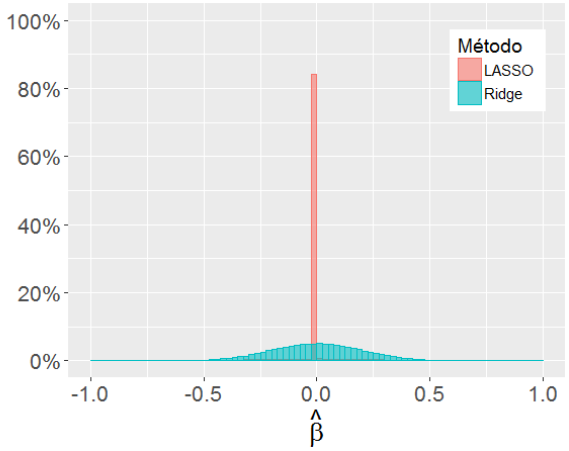
(g) $k = 20$



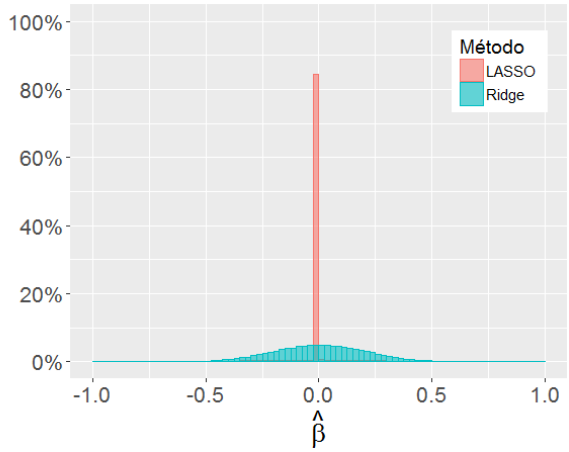
(h) $k = 25$



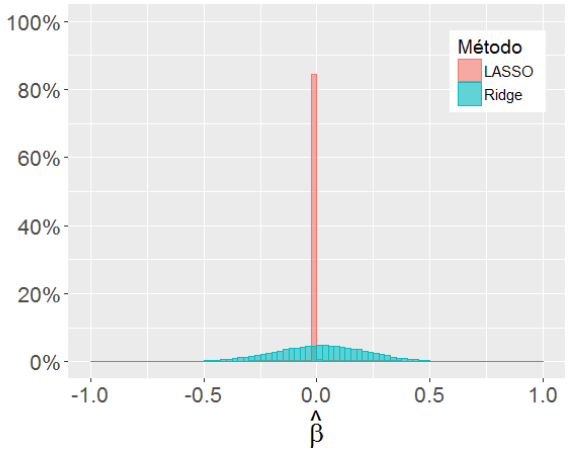
(i) $k = 30$



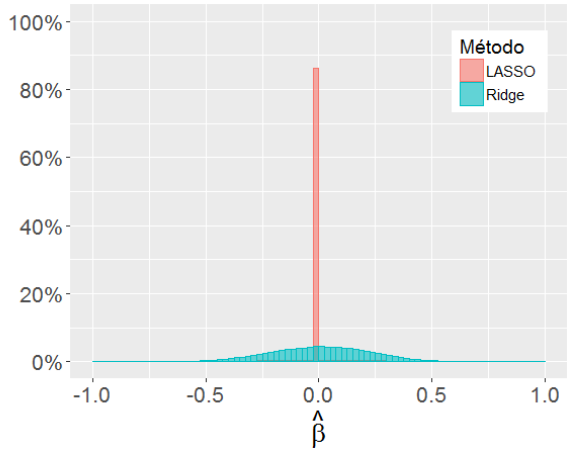
(j) $k = 35$



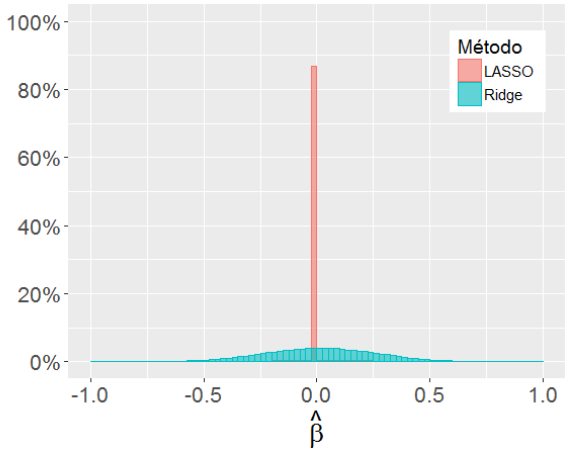
(k) $k = 40$



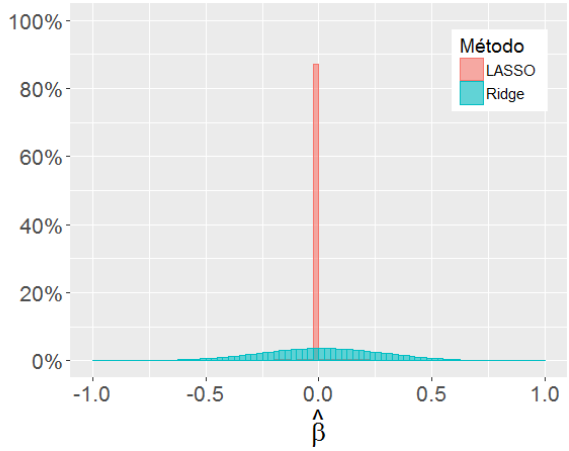
(l) $k = 45$



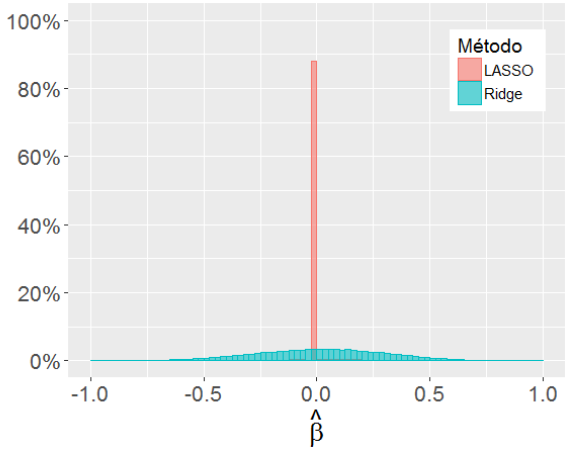
(m) $k = 50$



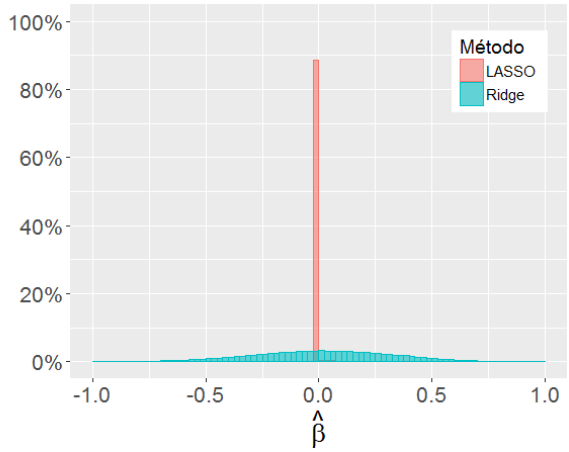
(n) $k = 60$



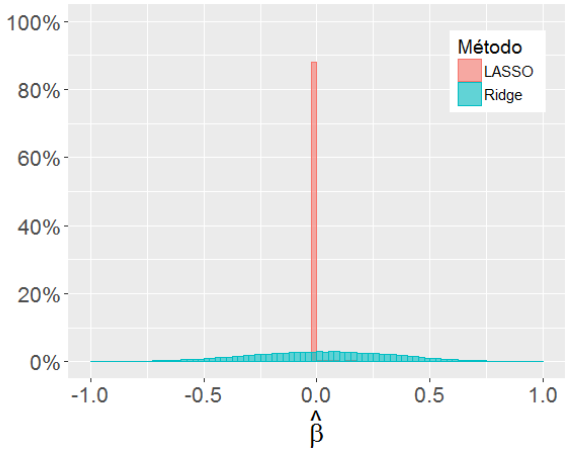
(ñ) $k = 70$



(o) $k = 80$

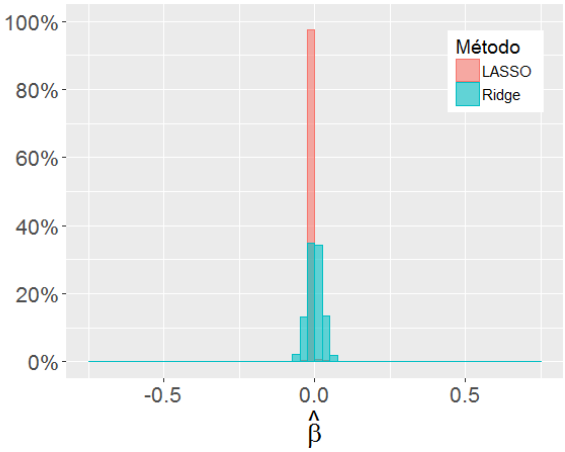


(p) $k = 90$

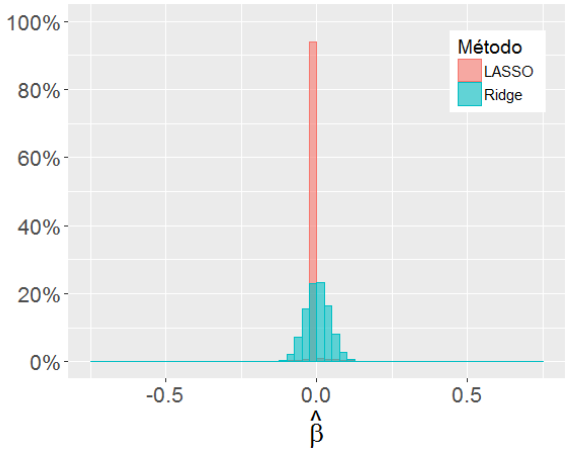


(q) $k = 100$

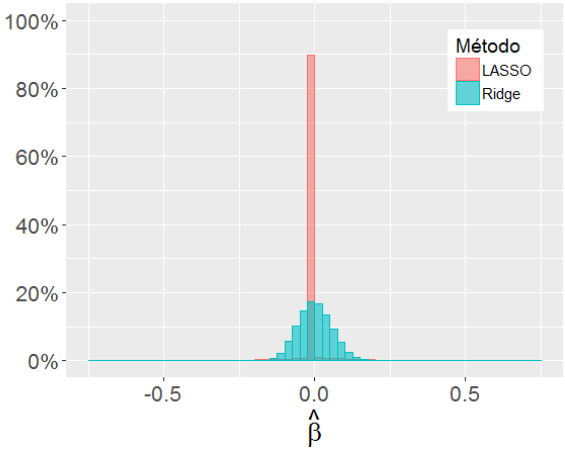
Caso $p = 400$.



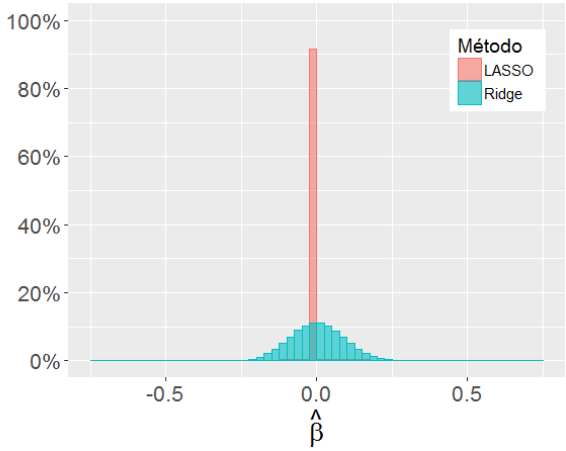
(a) $k = 2$



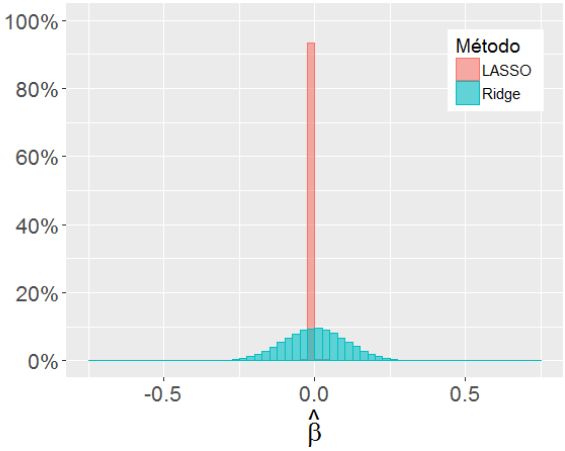
(b) $k = 5$



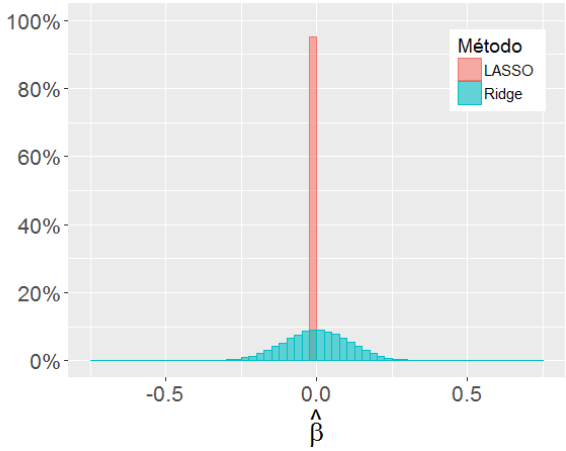
(c) $k = 10$



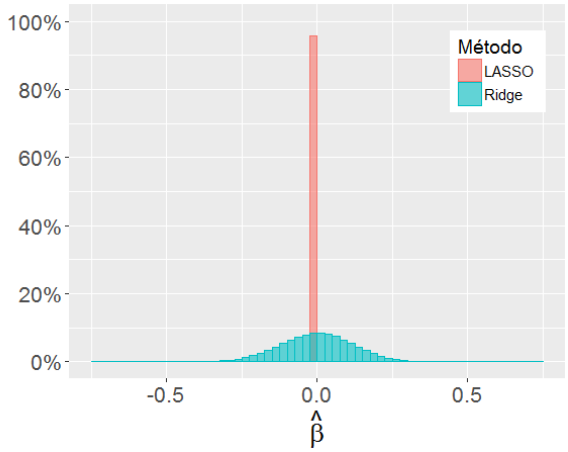
(d) $k = 20$



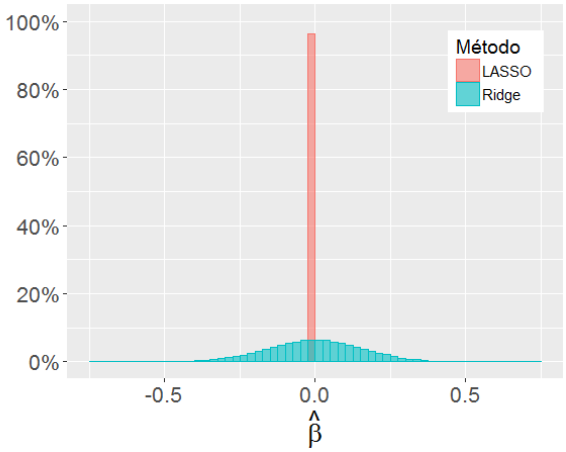
(e) $k = 30$



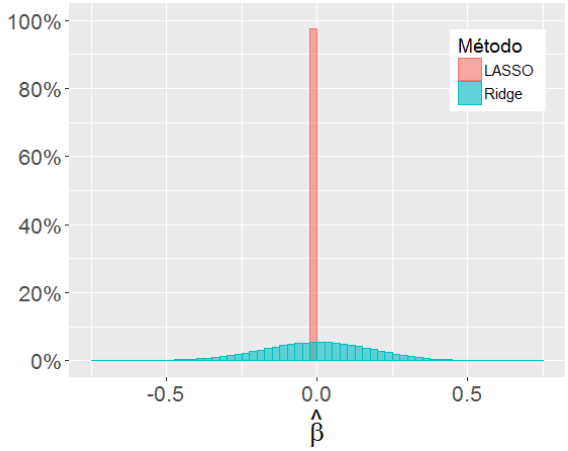
(f) $k = 40$



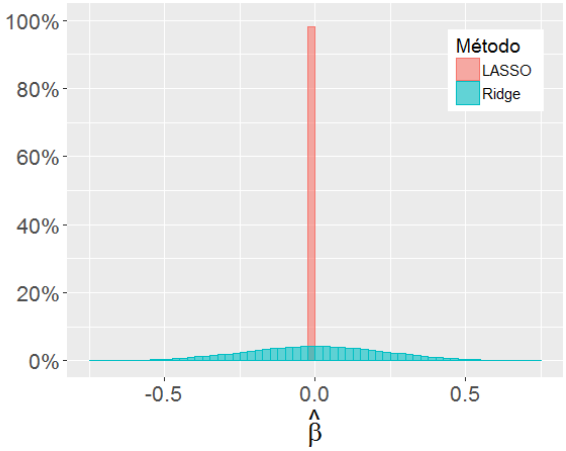
(g) $k = 50$



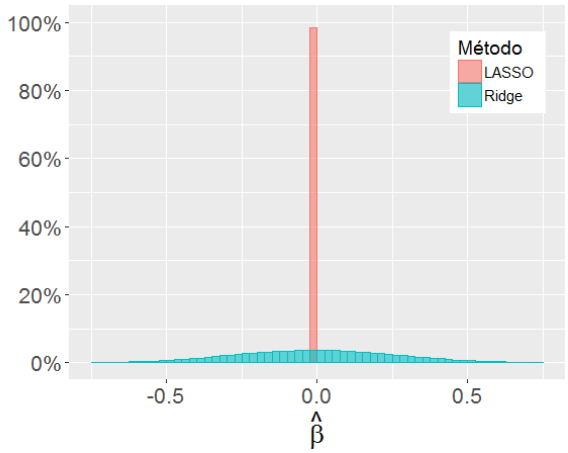
(h) $k = 75$



(i) $k = 100$



(j) $k = 150$



(k) $k = 200$