ANÁLISIS DE DATOS EN GRANDES DIMENSIONES. ESTIMACIÓN Y SELECCIÓN DE VARIABLES EN REGRESIÓN.

Sebastián Castro - scastro@iesta.edu.uy [1]

RESUMEN

Distintos avances tecnológicos han provocado un profundo impacto en la sociedad en general y en diversas áreas de la investigación científica en particular. A partir de algunos de estos avances es posible recolectar cantidades masivas de información respecto a ciertos fenómenos de interés a un costo relativamente bajo. La disponibilidad de estas enormes bases de datos y el objetivo de extraer información valiosa de ellas plantea nuevos desafíos para el análisis estadístico. Las técnicas de *selección de variables* y de *reducción de dimensiones* son fundamentales en este contexto debido a que modelos más *parsimoniosos* son deseables desde el punto de vista de la interpretación así como de la reducción en los errores de predicción. Por ejemplo, algunos métodos tradicionales de selección de variables en regresión, como AIC, BIC, C_p de Mallows y métodos secuenciales (*forward selection* y *bacward elimination*), pueden resultar fuertemente inestables o directamente inaplicables cuando el número de variables p es similiar o incluso ampliamente superior al número de observaciones n, conocido como *el caso* $p \gg n$. Debido a ésto, nuevas metodologías se han desarrollado en las últimas décadas que permiten enfrentar el problema o *maldición de la dimensionalidad*. Un conjunto amplio de estas técnicas puede plantearse agregando a la función objetivo, que mide el ajuste de los datos a un determinado modelo, un término de *penalización* o *regularización* por complejidad. El objetivo principal de este trabajo consiste en presentar algunos de estos desarrollos en el contexto de los modelos de regresión lineal y mostrar su aplicación e implementación sobre un conjunto de datos simulados con las características antes mencionadas. Al mismo tiempo se buscará indicar posibles aplicaciones y futuras líneas de investigación de especial interés en *econometría*.

Palabras clave: *selección de variables, regularización, Ridge, LASSO, SCAD.*

1. Introducción

En términos generales, el problema de selección de variables en regresión surge cuando se quiere modelar la relación entre una (o más) variable(s) de interés y un conjunto de potenciales variables explicativas $X_1, \ldots, X_p$, pero existe incertidumbre acerca de cuál subconjunto de las variables X_j utilizar. Esta situación es especialmente interesante y desafiante cuando p es *grande* y una buena parte de las variables $X_1, \ldots, X_p$ consideradas, son redundantes o irrelevantes (George, 2000). En general, si se incluyen cada vez más variables en un modelo de regresión, el ajuste a los datos de entrenamiento mejora, aumenta la cantidad de parámetros a estimar pero disminuye su precisión individual (mayor variancia) y por tanto la de la función de regresión estimada. A partir de cierto

[1]Instituto de Estadística (IESTA) y Departamento de Métodos Cuantitativos - Área Matemática, Facultad de Ciencias Económicas y de Administración, Universidad de la República.

punto empeora la capacidad de generalización del modelo (mayor error de predicción sobre datos nuevos), con lo cual se produce *sobreajuste*. En el otro extremo, si se incluyen muy pocas variables en el modelo, las variancias serán reducidas pero los sesgos mayores, obteniéndose una pobre descripción de los datos (*subajuste*). Algún tipo de compromiso entre estos dos escenarios extremos es por lo tanto deseable (Izenman, 2008).

Una de las características distintivas del problema de selección de variables es su enorme tamaño dado que, incluso para un número moderado de variables p, la evaluación de ciertas características de todos los 2^p modelos que permitan compararlos resulta excesivamente costosa o directamente imposible. Actualmente, la selección del mejor subconjunto sólo puede ser implementada con p cercano a 40 ($2^{40} > 10^{12}$) (Sheather, 2009; Hastie y otros, 2009). Por lo tanto, este no es un camino viable en muchos de los problemas actuales donde el número de potenciales variables predictoras es del orden de mil o más, con lo cual alguna reducción del *espacio de modelos* se vuelve necesaria.

Una posible reducción del espacio de modelos se puede obtener mediante distintas variantes de métodos de a pasos (*stepwise*), que secuencialmente agregan (*forward selection*) o eliminan (*bacward elimination*) variables de a una por vez entre la consideración de un modelo y el siguiente. Estos métodos son ejemplos de *algoritmos greedy*, donde la búsqueda de un óptimo global se reemplaza por la consideración sucesiva de óptimos locales. A su vez, tanto *forward selection* como *bacward elimination* consideran a lo sumo $p + (p-1) + \ldots + 1 = p(p+1)/2$ subconjuntos de los 2^p posibles, y por lo tanto no necesariamente encuentran el modelo óptimo global (además de que tampoco es seguro que ambos métodos elijan el mismo modelo final). Otra desventaja importante de los métodos secuenciales es su *inestabilidad*, en el sentido de que pequeños cambios en el conjunto de datos pueden producir grandes modificaciones en los resultados, en particular en las variables seleccionadas (Breiman, 1996). A su vez, todos estos inconvenientes se ven agravados en presencia de predictores fuertemente correlacionados, debido a que una variable con efecto verdadero sobre la respuesta puede no ser elegida si alguna otra variable correlacionada con ésta ya ha ingresado al modelo.

Una vez que el conjunto de modelos a considerar se ha reducido a un tamaño *manejable*, se necesita un criterio para poder evaluarlos y compararlos. Algunas posibles medidas están vinculadas al nivel de *ajuste* de cada modelo (aunque es deseable que incluyan también una penalización por *complejidad*) o al error de *predicción* del modelo sobre nuevas observaciones. En el contexto de modelos lineales, en el primer tipo de medidas se encuentran criterios clásicos como el R^2-ajustado y *Criterios de Información* (AIC, BIC o GIC), mientras que en el segundo se ubican, por ejemplo, el estadístico C_p de *Mallows* o la técnica de *validación cruzada*.

2. Técnicas de regularización en modelos lineales

Debido a que las técnicas clásicas de selección de variables realizan un *proceso discreto* de exploración del espacio de modelos (cada variable es seleccionada o descartada), éstas sufren frecuentemente de alta variabilidad lo cual a su vez puede perjudicar su desempeño en términos de predicción. En cambio, las técnicas de regularización o *shrinkage* son procedimientos *más continuos* y *menos variables*, con lo cual se presentan como alternativas interesantes (Hastie y otros, 2009). Los métodos de regularización pueden aplicarse a una

amplia variedad de modelos, no solamente modelos lineales, aunque es sobre estos últimos donde más desarrollo se ha hecho y sobre los cuales se tratará en este trabajo. En términos generales, los métodos de regularización buscan convertir un problema *mal condicionado*, debido a la no unicidad de la solución, en uno *bien condicionado*. Este resulta ser entonces un marco adecuado para trabajar los problemas de regresión en el caso $p \gg n$.

En términos generales, la regularización permite entrenar modelos complejos con conjuntos de datos relativamente pequeños, disminuyendo el riesgo de sobreajuste mediante el control de la complejidad del modelo. En el contexto de los modelos lineales, el problema de determinar la complejidad óptima del modelo se traslada de encontrar el número apropiado de funciones base (proceso discreto) a uno en el cual debe determinarse un valor adecuado del parámetro de *regularización* λ (proceso continuo). Una formulación amplia de las técnicas de regularización en el contexto de modelos lineales (con variable de respuesta continua) puede realizarse de la siguiente manera:

$$\hat{\beta} = \arg\min_{\beta_0,\beta} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \sum_{j=1}^{p} \phi_\lambda(|\beta_j|) \right\} \tag{1}$$

donde $\beta = (\beta_1, \ldots, \beta_p)$, $\lambda \geq 0$ y $\phi_\lambda(|\beta_j|)$ es una función creciente de penalización sobre el "tamaño" de β, que depende a su vez de λ.

Una familia de funciones de penalización muy utilizada es la correspondiente a la norma-L_q, dada por $\sum_{j=1}^{p} \phi_\lambda(|\beta_j|) = (\|\beta\|_q)^q = \sum_{j=1}^{p} |\beta_j|^q$, para $q > 0$ (aunque estrictamente solo podamos hablar de norma cuando $q \geq 1$). Los estimadores resultantes en este caso son conocidos como *estimadores Bridge* (Fu, 1998) y en especial los casos $0 < q \leq 1$ se conocen como de *umbralización suave* (*soft thresholding* en inglés). Los métodos de selección de modelos que penalizan por el número de parámetros (AIC, BIC, R^2_{ajust}, etc.) pueden ser vistos a su vez como casos límites de estimadores Bridge cuando $q \to 0$, dado que en ese caso $|\beta_j|^q \to 0$ si $\beta_j = 0$ y $|\beta_j|^q \to 1$ cuando $\beta_j \neq 0$.

Una formulación alternativa de (1) corresponde a resolver:

$$\min_{\beta_0,\beta} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \right\}, \text{ sujeto a } \sum_{j=1}^{p} \phi_\lambda(|\beta_j|) \leq s \tag{2}$$

donde $s \geq 0$ es un parámetro de ajuste (*tunning*). En esencia, (1) es la *forma lagrangiana* del problema de optimización con restricciones (2). Por lo tanto, ambos problemas son equivalentes en el sentido de que si $\hat{\beta}_\lambda$ es la solución de (1) y $\hat{\beta}_s$ es la solución de (2), entonces para cada $\lambda_0 > 0$ y la solución correspondiente $\hat{\beta}_{\lambda_0}$, existe s_{λ_0} tal que $\hat{\beta}_{\lambda_0} = \hat{\beta}_{s_{\lambda_0}}$. Y vicerversa, dado cualquier $s_0 > 0$ y el correspondiente $\hat{\beta}_{s_0}$, existe λ_{s_0} tal que $\hat{\beta}_{s_0} = \hat{\beta}_{\lambda_{s_0}}$. Es decir, existe una *correspondencia uno a uno* entre λ y s (Clarke y otros, 2009).

En el caso de una penalización mediante norma-L_q, la solución al problema de regresión lineal (regularizado) se expresa como:

$$\hat{\beta} = \arg\min_{\beta_0,\beta} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j|^q \right\} \tag{3}$$

En las Figura 1 se puede visualizar las curvas de nivel de esta función de penalización para el caso de dos variables: $\phi(\beta_1, \beta_2) = |\beta_1|^q + |\beta_2|^q$. Se observa que solamente para $q \geq 1$ la penalización es *convexa* y, por lo tanto, también es convexo el conjunto factible del problema de optimización con restricciones (2). La convexidad de un problema de optimización es una característica deseable desde el punto de vista computacional.

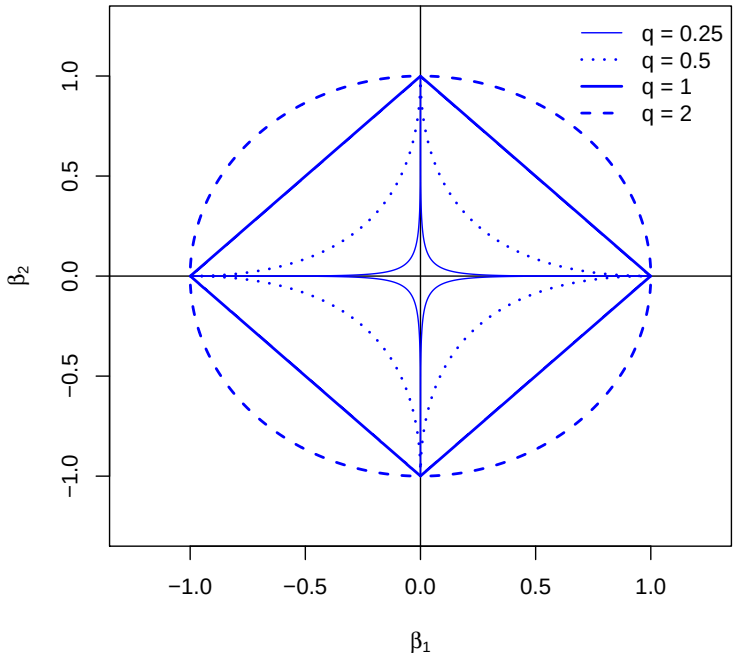

Figura 1: Curvas de nivel de la penalización L_q en dos dimensiones.

Para evitar que la penalización varíe frente a cambios de escala de las variables, habitualmente éstas son *estandarizadas* (media 0 y variancia 1) previamente. De esta forma, se puede ajustar un modelo sin término independiente estimando éste mediante $\bar{y}_n$ (Hastie y otros, 2009). Notar además que el término β_0 no es incluido en la penalización de forma de evitar que el resultado dependa del origen en la variable y.

Como puede observarse de (1), todas estas técnicas dependen de un parámetro de *complejidad* λ, que controla la importancia dada a la penalización en el proceso de optimización. Cuanto mayor es el parámetro de complejidad mayor es la penalización en los coeficientes de regresión y más son contraídos éstos hacia cero (*shrinkage*). En los casos extremos, si λ es igual a 0 la estimación (3) coincide con la de mínimos cuadrados

(MCO) habitual (cuando ésta es única), mientras que si $\lambda \to \infty$ entonces $\hat{\beta} \to 0 \in R^p$. El objetivo en cada problema particular será encontrar un valor adecuado $0 < \lambda < \infty$, lo cual en la práctica suele hacerse mediante *validación cruzada* o *bootsrap*, con el propósito de minimizar una estimación del error de predicción esperado.

Tanto *Ridge* como *LASSO*, técnicas que veremos en las siguientes secciones, representan casos particulares de (3) y ambas plantean la estimación de un modelo de regresión lineal pero difieren en la forma de penalización del vector de parámetros (norma-L_2 y norma-L_1 respectivamente). Esta diferencia en la penalización puede parecer marginal pero tiene grandes consecuencias. El uso de la norma-L_2 tiene el efecto agradable de producir un estimador lineal en y del vector de parámetros β pero como contrapartida utiliza todas las variables predictoras en el modelo de regresión final, debido a que valores mayores de λ contraen los coeficientes hacia cero pero sin alcanzar dicho valor en general. Por su parte LASSO, mediante la penalización L_1, no produce un estimador lineal en y ni se obtiene una fórmula cerrada para su expresión sino que debe encontrarse la solución a través de un algoritmo de optimización. Sin embargo, como se verá más adelante dependiendo de la elección del parámetro de complejidad la penalización L_1 produce algunos coeficientes de regresión exactamente nulos. Esto tiene la ventaja de que en el modelo final solamente algunas de las variables son consideradas, presentándose entonces como un método de estimación y de selección de variables al mismo tiempo.

Todas estas técnicas son conocidas como métodos de *regularización* o *shrinkage* porque contraen los coeficientes de regresión con el objetivo de estabilizar la estimación. Esta regularización implica que el *tamaño* del vector de parámetros es restringido a cierto rango, evitando de esta forma que variables explicativas altamente correlacionadas produzcan estimaciones mínimo cuadráticas fuertemente inestables o simplemente permitiendo que se produzcan estimaciones únicas (cuando existe multicolinealidad o el número de variables supera la cantidad de observaciones). Estos métodos son típicamente utilizados para regresión de una variable dependiente y sobre una matriz de altas dimensiones X, con variables muy correlacionadas. Además de la *Genética* y la *Bioinformática* (Li y Xu, 2009), otras áreas de actual aplicación son el *Procesamiento de Señales*, la *Quimiometría* (Varmuza y Filzmoser, 2009) y la *Econometría* (Belloni y Chernozhukov, 2011; Belloni, Chernozhukov y Hansen, 2011; Fan, Lv y Qi, 2011).

2.1. Regresión Ridge

Esta técnica fue propuesta originalmente en 1970 como un método para lidiar con el problema de *colinealidad* en un modelo lineal estimado por mínimos cuadrados, en el contexto $p < n$ (Hoerl y Kennard, 1970). Recordemos que cuando existen predictores altamente correlacionados en el modelo, la estimación de los coeficientes resulta ser muy inestable (variancia grande). Posteriormente, Regresión Ridge se incorporó a la clase más amplia de técnicas de regularización en la forma en que han sido presentadas anteriormente.

Recordando que $\hat{\beta}^{mco} = (\mathrm{X}^{tr}\mathrm{X})^{-1}\mathrm{X}^{tr}y$ es la estimación por mínimos cuadrados de β, se planteó en un principio que la potencial inestabilidad de $\hat{\beta}^{mco}$ podría ser aliviada agregando una pequeña constante $k \geq 0$ a cada término de la diagonal de $\mathrm{X}^{tr}\mathrm{X}$ antes de invertir la matriz (Hoerl y Kennard, 1970). Este proceso resulta en el estimador Ridge:

$$\hat{\beta}^{ridge} = (\mathrm{X}^{tr}\mathrm{X} + k\mathrm{I}_p)^{-1}\mathrm{X}^{tr}y \tag{4}$$

siendo I_p la matriz identidad de dimensión $p \times p$. Más precisamente, (4) representa una familia de estimadores, un estimador para cada valor de k.

El principal problema a resolver entonces en la aplicación de Regresión Ridge es la determinación del valor de k más adecuado. En forma interesante, la elección de k involucra un balance entre los componentes de sesgo y variancia del error cuadrático medio al estimar β. En este sentido (y asumiendo un modelo lineal), cuanto mayor es k más grande es el sesgo pero menor es la variancia y la determinación final implica un compromiso entre ambos términos (Izenman, 2008). En general, Regresión Ridge produce predicciones más precisas que los modelos obtenidos por mínimos cuadrados más selección "clásica" de variables, a menos que el verdadero modelo sea *esparsa* (mayoría de coeficientes nulos). Una propuesta inicial y que continúa siendo sugerida por algunos autores es la utilización de una *traza ridge* para determinar k. La traza ridge es un gráfico simultáneo de los coeficientes de regresión estimados (4) (los cuales dependen de k) respecto del parámetro. Luego, el valor de k se elige como el menor de todos los considerados para los cuales se estabilizan los coeficientes estimados. Este método presenta cierto grado de arbitrariedad como forma de elegir el modelo final y a menudo más que un único valor de k sugiere un rango de valores adecuados. Otros métodos más automáticos consisten en estimar el parámetro mediante *validación cruzada* o *bootsrap*. En general se recomienda utilizar todos los métodos y comparar los resultados.

Como se mencionó anteriormente, Regresión Ridge puede verse como un caso particular de las técnicas de regularización restringiendo la norma L_2 de los coeficientes del modelo. Es decir, se puede obtener $\hat{\beta}^{ridge}$ como solución del problema:

$$\min_{\beta_0,\beta} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \right\}, \text{ sujeto a } \sum_{j=1}^{p} \beta_j^2 \leq s \tag{5}$$

O en forma equivalente:

$$\hat{\beta}^{ridge} = \arg\min_{\beta_0,\beta} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\} \tag{6}$$

Una interpretación gráfica útil del proceso de optimización sujeto a restricciones (5) puede verse en la Figura 2 para el caso de dos dimensiones. Allí se muestran las curvas de nivel de $SCR(\beta) = ||y - \mathrm{X}\beta||_2^2$ junto con la región factible $||\beta||_2^2 \leq s$. Se observa cómo la estimación ridge *contrae* los coeficientes $\hat{\beta}_j$ hacia 0 respecto de los obtenidos mediante MCO.

2.2. Regresión LASSO

LASSO (*Least Absolute Shrinkage and Selection Operator*), introducida en la comunidad estadística en 1996 (Tibshirani, 1996), es una técnica de regresión lineal regularizada

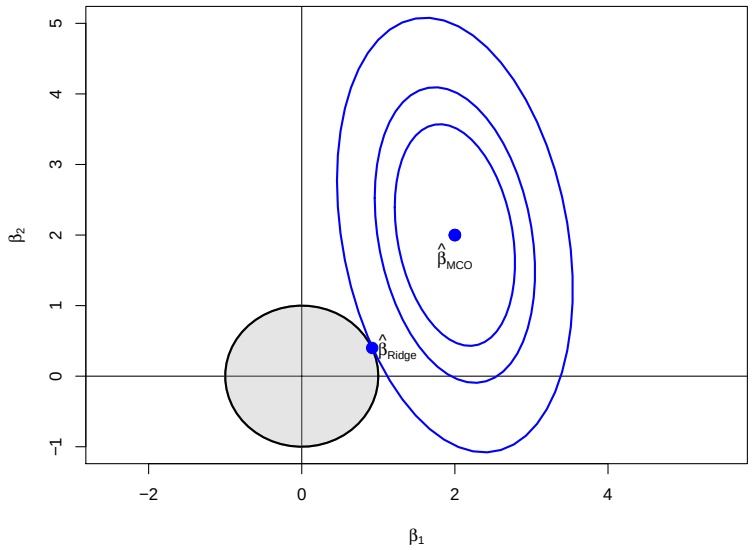

Figura 2: Descripción gráfica de la estimación Ridge en dos dimensiones.

como Ridge, con una leve diferencia en la penalización que trae consecuencias importantes. En especial, a partir de cierto valor del parámetro de complejidad el estimador de LASSO produce estimaciones nulas para algunos coeficientes y no nulas para otros (soluciones *esparsas*), con lo cual LASSO realiza una especie de selección de variables en forma continua. Esto no sucede con Ridge, donde por lo general todos los coeficientes son contraídos al mismo tiempo hacia cero sin llegar a alcanzar este valor. La motivación inicial para LASSO fue entonces tener una técnica que mediante la contracción de los coeficientes lograra estabilizar las estimaciones y predicciones (como Ridge) pero que a su vez produjera modelos más estables e interpretables mediante la selección de variables (Tibshirani, 1996). Sin embargo, el auge en la investigación de técnicas tipo LASSO es bastante reciente debido a la abundancia actual de problemas que pueden expresarse como regresión en el caso $p \gg n$, y la facilidad y disponibilidad computacional (Tibshirani, 2011).

El estimador de LASSO, $\hat{\beta}^{LASSO}$, se define como solución del problema de optimización con restricciones:

$$\min_{\beta_0,\beta} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \right\}, \text{ sujeto a } \sum_{j=1}^{p} |\beta_j| \leq s \tag{7}$$

O mediante el equivalente lagrangiano como:

$$\hat{\beta}^{LASSO} = \arg\min_{\beta_0,\beta} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\} \tag{8}$$

La penalización mediante la norma L_1 produce un *estimador no lineal* en la variable de respuesta y, y no existe en general una expresión "en forma cerrada" de $\hat{\beta}^{LASSO}$, a diferencia de MCO y Ridge. El cálculo del estimador LASSO para un valor dado de λ se

puede obtener reescribiendo (7) como un *problema de programación cuadrática*, donde se busca minimizar una función objetivo cuadrática sujeto a restricciones lineales en las variables β_j. Sin embargo, como veremos más adelante existen varios algoritmos eficientes que permiten obtener la solución para cada valor de λ con el mismo costo computacional que en Ridge. Al igual que en todas las técnicas de penalización consideradas, el parámetro de regularización λ debería ser elegido en función de los datos con el propósito de minimizar una estimación del error de predicción esperado. Nuevamente, *validación cruzada* y *bootsrap* son las alternativas preferidas en general. Adicionalmente, el gráfico de los coeficientes estimados en función de λ es una herramienta útil para visualizar la evolución del ajuste a medida que aumenta la penalización.

En la Figura 3 se muestra el proceso de estimación para el caso de dos variables. La solución se establece en el primer punto donde los contornos elípticos se encuentran con la región factible, representado por el cuadrado $|\beta_1| + |\beta_2| \leq s$. A diferencia de la restricción en Ridge, la solución ocurre habitualmente en los vértices del cuadrado donde alguno de los β_j es igual a cero. Cuando $p > 2$, el cuadrado se convierte en un hipercubo con mayor cantidad de vértices y por lo tanto con mayor oportunidad para que los parámetros estimados sean nulos (Hastie y otros, 2009).

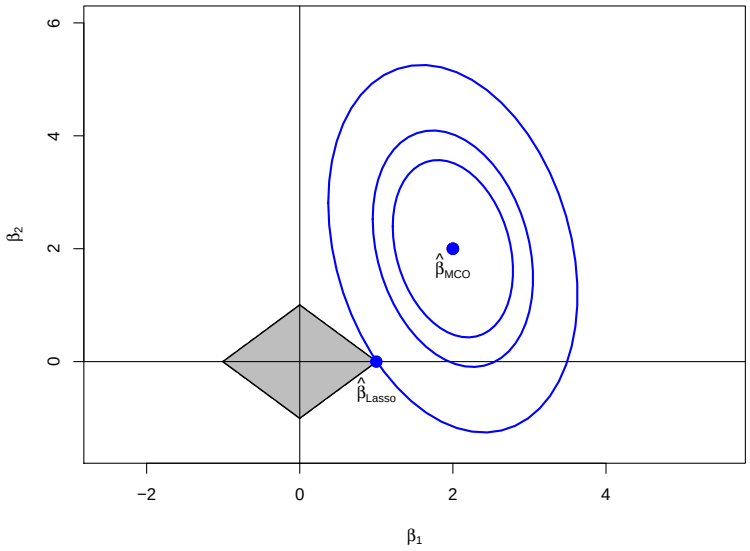

Figura 3: Descripción gráfica de la estimación LASSO en dos dimensiones.

2.2.1. Alternativas para la estimación

Como se mencionó anteriormente, la obtención del estimador de Lasso plantea desafíos adicionales desde el punto de vista computacional. Una de las primeras alternativas consistió en escribir el problema original (7) como un *problema de programación cuadrática*, observando que $\beta_j = \beta_j^+ - \beta_j^-$ y $|\beta_j| = \beta_j^+ + \beta_j^-$, donde $\beta_j^+ = \text{máx}\{0, \beta_j\}$ y $\beta_j^- = \text{máx}\{0, -\beta_j\}$. Luego, se resuelve el problema de mínimos cuadrados en las nuevas variables β_j^+ y β_j^-, con las restricciones lineales $\sum_{j=1}^p \beta_j^+ + \sum_{j=1}^p \beta_j^- \leq s$, y $\beta_j^+, \beta_j^- \geq 0$. De

esta forma, para un valor dado de s o λ se transforma el problema de optimización original (de p variables y 2^p restricciones) en un nuevo problema equivalente con más variables ($2p$) pero menos restricciones ($2p + 1$), que puede ser resuelto por técnicas estándar de programación cuadrática (Tibshirani, 1996).

Sin embargo, posteriormente se propusieron alternativas más eficientes que permiten obtener el camino entero de soluciones $\hat{\beta}^{LASSO}(\lambda)$ en un sólo paso, es decir, calcular la solución de Lasso simultáneamente para todos los valores de λ. Efron y otros (2004), mostraron que Lasso tiene un camino de solución lineal por tramos. Esto implica que para algún m natural existen $\lambda_0 = 0 < \lambda_1 < \ldots < \lambda_m = \infty$ y $\xi_0, \ldots, \xi_{m-1} \in R^p$ tales que:

$$\hat{\beta}^{LASSO}(\lambda) = \hat{\beta}^{LASSO}(\lambda_k) + (\lambda - \lambda_k)\xi_k \text{ con } \lambda_k \leq \lambda \leq \lambda_{k+1}, \tag{9}$$

para $k = 0, \ldots, m-1$. De esta forma, es posible generar el camino completo de solución $\hat{\beta}^{LASSO}(\lambda), 0 \leq \lambda \leq \infty$, calculando secuencialmente el tamaño del paso entre los valores de λ y las direcciones $\xi_0, \ldots, \xi_{m-1}$ (Clarke y otros, 2009). Uno de los algoritmos que se aprovecha de estos resultados es LARS (Least Angle Regression), recientemente propuesto como un algoritmo de ajuste y de selección de variables para modelos lineales (Efron y otros, 2004) con tres importantes características: *(i)* una simple modificación del algoritmo LARS produce el estimador de Lasso, *(ii)* una modificación diferente del algoritmo implementa otra técnica de selección de variables (*Forward Stagewise*) y *(iii)* se obtiene una aproximación simple de los grados de libertad del estimador LARS que permite derivar una estimación del error de predicción (Efron y otros, 2004; Hastie y otros 2009).

Un método alternativo al algoritmo LARS para la estimación de LASSO es el de *Coordenada Descendente* (Hastie y otros, 2009). La idea principal consiste en fijar el parámetro de penalización λ y optimizar sucesivamente respecto de cada parámetro β_j, dejando los restantes parámetros β_k, $k \neq j$, fijos en sus valores actuales. De acuerdo a estudios realizados sobre datos simulados y reales, éste método puede ser más rápido (menor tiempo computacional) que LARS, especialmente cuando $p \gg n$ (Friedman y otros, 2010). Por otro lado, mediante los algoritmos de coordenada descendente se obtienen estimaciones sobre una *grilla* de valores de λ y no el camino completo de soluciones (como ocurre en LARS). Adicionalmente, existen extensiones e implementaciones computacionales de este tipo de algoritmos a *Modelos Lineales Generalizados* con penalización L_1 (Friedman y otros, 2010).

2.3. Penalizaciones no convexas (SCAD)

En los últimos años se han desarrollado algunas generalizaciones y extensiones de las técnicas presentadas anteriormente, especialmente diseñadas para ciertas situaciones particulares donde Ridge, LASSO y en general la penalización L_q, podrían no ser satisfactorias. Todas ellas buscan retener las ventajas de LASSO como método de estimación y selección de variables, y al mismo tiempo corregir algunas de sus posibles desventajas.

Como una variante de los métodos de penalización de la forma (3) (estimadores Bridge), Zou y Hastie (2005), propusieron Elastic Net, un método de penalización que representa un compromiso entre las penalizaciones L_1 y L_2. Otras variantes están dadas

por LASSO Adaptativo (Zou, 2006) y LASSO Relajado (Meinshausen, 2006). Todas estas técnicas tienen la ventaja de utilizar penalizaciones convexas con lo cual se aprovechan de la existencia de algoritmos eficientes para su implementación.

Por otro lado, en un reciente trabajo Fan y Li (2001) propusieron tres condiciones deseables que un método de penalización debería cumplir:

1. *esparsidad*; efectuar selección de variables automáticamente, estableciendo que coeficientes suficientemente pequeños sean nulos.

2. *continuidad*; ser continuo en los datos para evitar inestabilidad en la predicción.

3. *insesgadez*; tener bajo sesgo, especialmente para valores grandes de los coeficientes β_j.

Las técnicas de penalización L_q, $0 \leq q < 1$, no satisfacen la condición de continuidad, la penalización L_1 (LASSO) no satisface la condición de insesgadez y L_q, $q > 1$ (Ridge), no verifica la condición de esparsidad. Por lo tanto, ninguna de las técnicas de penalización L_q satisfacen las tres condiciones simultáneamente (Fan y Li, 2001).

Como alternativa, estos autores proponen la penalización SCAD (*Smoothly Clipped Absolute Deviation*):

$$\phi_\lambda(\beta_j) = \begin{cases} \lambda|\beta_j| & \text{si } 0 \leq |\beta_j| \leq \lambda \\ -(\beta_j^2 - 2a\lambda|\beta_j| + \lambda^2)/(2(a-1)) & \text{si } \lambda \leq |\beta_j| \leq a\lambda \\ (a+1)\lambda^2/2 & \text{si } |\beta_j| \geq a\lambda \end{cases} \tag{10}$$

donde $a > 2$ y $\lambda > 0$ son parámetros de ajuste. La penalización SCAD es muy similar a L_1 (Lasso) para valores pequeños de β_j, mientras que para valores grandes la primera es constante y la última no. Esto muestra la diferencia entre ambas en la propiedad de insesgadez (ver Figura 4).

El estimador de SCAD, $\hat{\beta}^{SCAD}$, se define para a y λ fijos, como el que minimiza:

$$\sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \sum_{j=1}^{p} \phi_\lambda(\beta_j) \tag{11}$$

Los parámetros a y λ pueden ser elegidos mediante validación cruzada aunque se recomienda utilizar $a \approx 3.7$ como valor por defecto para reducir el costo computacional (Fan y Li, 2001). El mayor desafío se encuentra en la implementación de SCAD, dado que se trata de un *problema no convexo*. Algunos de los algoritmos propuestos plantean realizar aproximaciones (convexas) locales de la función objetivo (Fan y Li, 2001; Clarke y otros, 2009; Fan y Lv, 2010) y utilizar iterativamente los algoritmos eficientes para penalizaciones convexas.

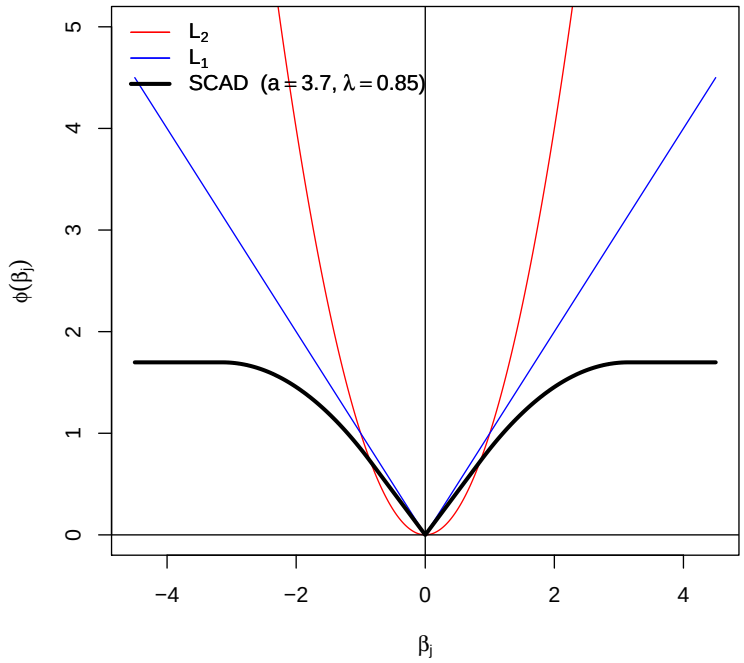

Figura 4: Penalizaciones en Ridge, LASSO y SCAD.

3. Extensión a modelos lineales generalizados

Las técnicas de penalización en regresión pueden extenderse a una amplia variedad de tipos de variable respuesta, incluyendo respuestas binarias, de conteo y continuas. Como se mencionó anteriormente, una familia popular de modelos en este contexto es el de los Modelos Lineales Generalizados, donde la variable de respuesta pertenece a la familia exponencial. Algunos de los casos más conocidos son los modelos de *regresión logística*, *multinomial*, *poisson*, *gamma*, *binomial negativa* y *normal/gaussiana* (Fan y Li, 2001; Fan y Li, 2006; Friedman y otros, 2010).

Supongamos que dado $\mathbf{x}_i = (x_1, \ldots, x_p)$, Y_i tiene densidad $f(y_i|g(\mathbf{x}_i^{tr}\beta))$, donde g es una *función de enlace* conocida y $\log f$ denota la log-verosimilitud condicional de Y_i. Luego, se define la verosimilitud penalizada como:

$$\sum_{i=1}^{n} \log f(y_i|g(\mathbf{x}_i^{tr}\beta)) - n\sum_{j=1}^{p} \phi_\lambda(\beta_j) \tag{12}$$

Maximizar la verosimilitud penalizada respecto de β es equivalente a minimizar:

$$-\sum_{i=1}^{n} \log f((y_i|g(\mathbf{x}_i^{tr}\beta)) + n\sum_{j=1}^{p} \phi_\lambda(\beta_j) \tag{13}$$

lo cual generaliza lo presentado hasta ahora sobre respuestas continuas.

4. Una *perspectiva bayesiana* sobre las técnicas de regularización

Bajo *distribuciones a priori* no informativas estándar, el análisis bayesiano del modelo de regresión lineal (para $p < n$) tiene varios puntos en común con los resultados obtenidos por MCO y máxima verosimilitud.

Por ejemplo, partiendo del modelo $y|\beta, \sigma^2, \mathrm{X} \sim N(\mathrm{X}\beta, \sigma^2 \mathrm{I}_n)$ para los datos y la distribución a priori no informativa $p(\beta, \sigma^2|\mathrm{X}) \propto \sigma^{-2}$, tenemos:

$$p(y|\beta, \sigma^2, \mathrm{X}) \propto \exp\left\{-\frac{1}{2\sigma^2}(y - \mathrm{X}\beta)^{tr}(y - \mathrm{X}\beta)\right\}$$
$$\propto \exp\left\{-\frac{1}{2\sigma^2}(\beta - \hat{\beta}^{mco})^{tr}\mathrm{X}^{tr}\mathrm{X}(\beta - \hat{\beta}^{mco})\right\}$$

Con lo cual, la distribución (condicional) *a posteriori* de β es:

$$\beta|y, \sigma^2, \mathrm{X} \sim N(\hat{\beta}^{mco}, \sigma^2(\mathrm{X}^{tr}\mathrm{X})^{-1}) \tag{14}$$

Mientras que la distribución (marginal) *a posteriori* de σ^2 resulta:

$$\sigma^2|y, \mathrm{X} \sim Inv - \chi^2(n - p, s^2)$$
$$\text{con } s^2 = (y - \mathrm{X}\hat{\beta}^{mco})^{tr}(y - \mathrm{X}\hat{\beta}^{mco})/(n - p)$$

El estimador $\hat{\beta}^{mco}$ es entonces la media, modo y mediana (condicional) a posteriori de β, bajo a prioris no informativas.

Utilizando distintas *distribuciones a priori informativas*, varias de las técnicas de regularización presentadas pueden ser vistas como estimadores bayesianos. Por ejemplo, si a priori $\beta|\sigma^2_\beta \sim N(0, \sigma^2_\beta \mathrm{I}_p)$, definiendo $\lambda = \sigma^2/\sigma^2_\beta$, se obtiene:

$$\beta|y, \sigma^2, \mathrm{X} \sim N(\hat{\beta}^{ridge}, \sigma^2(\mathrm{X}^{tr}\mathrm{X} + \lambda \mathrm{I}_p)^{-1}) \tag{15}$$

En cambio, tomando $p(\beta|\lambda) = \prod_{j=1}^{p} p(\beta_j|\lambda)$, con:

$$p(\beta_j|\lambda) = \frac{\lambda}{2}\exp\left\{-\lambda|\beta_j|\right\}, j = 1, \ldots, p$$

(distribución de Laplace o Doble Exponencial), se obtiene:

$$-2\log p(\beta|y, \lambda, \mathrm{X}) = (y - \mathrm{X}\beta)^{tr}(y - \mathrm{X}\beta) + \lambda\sum_{j=1}^{p}|\beta_j| + cte \tag{16}$$

Con lo cual $\hat{\beta}^{LASSO}$ coincide con el estimador *máximo a posteriori* (MAP) bajo este modelo. La diferencia en las distribuciones a priori consideradas puede verse en la Figura 5, donde se observa que la distribución *doble exponencial* concentra relativamente mayor

probabilidad cerca del origen. Otras técnicas de regularización pueden presentarse de esta manera donde la penalización se corresponde con una distribución a priori adecuada. Bajo el enfoque bayesiano, la complejidad en la estimación que implica resolver problemas de optimización en varias variables (convexos o no), se traslada al problema de simular de distribuciones multivariadas desconocidas a través de técnicas de *Cadenas de Markov de Monte Carlo* (MCMC) (Park y Casella, 2008; Hans, 2009 y 2010; Kyung y otros, 2010; Li y Lin, 2010; Celeux y otros, 2012).

Por último, bajo este marco parece más natural plantearse la pregunta de por qué utilizar Ridge, LASSO o alguna otra técnica frente a un problema dado. El conocimiento que se posee acerca del problema es fundamental para guiar la búsqueda de las herramientas más adecuadas. La *esparsidad* del modelo es en definitiva una a priori.

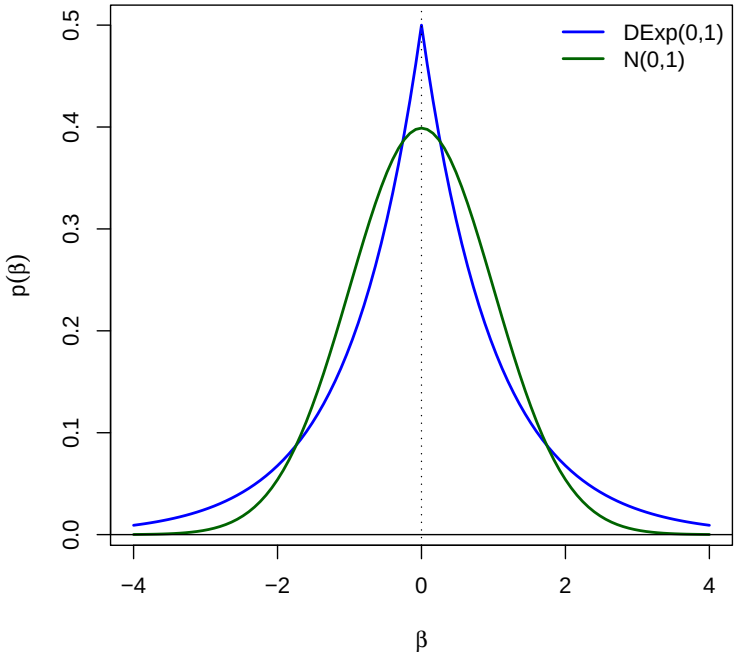

Figura 5: Ejemplos de distribuciones a priori implícitas en Ridge y LASSO.

5. Simulación

En esta sección el objetivo consiste en realizar un breve estudio de simulación para mostrar en la práctica la implementación de algunas de las técnicas presentadas. Para ello nos situamos en un contexto similar a los que han llevado al desarrollo de las mismas. En particular, se simulan covariables de altas dimensiones ($p \gg n$) y fuertemente correlacionadas entre sí, mientras que la variable de respuesta es generada a través de un modelo lineal *esparsa* (relativamente pocas variables con verdadero efecto). El hecho de analizar los resultados de las distintas técnicas sobre datos simulados tiene la ventaja de que se

conoce el *verdadero modelo* de asociación, con lo cual es posible evaluar en qué medida cada técnica recupera o descubre este modelo a través de una muestra de entrenamiento. Es importante observar de antemano que las características de los datos simulados pueden favorecer el desempeño de algunas técnicas frente a otras, con lo cual no es el objetivo concluir acerca de la conveniencia de uno u otro método en general. Tanto la simulación como el ajuste de las técnicas fue realizado utilizando el software libre R.

En concreto entonces, se simulan $n = 100$ observaciones de un modelo lineal con $p = 5000$ variables ($p \gg n$). En primera instancia se simula la matriz de predictores $X = ((x_{ij}))$, donde $x_{ij} \sim N(0,1)$, $\text{cor}(x_{\cdot j}, x_{\cdot k}) = \rho^{|j-k|}$ y $\rho = 0.85$, para $i = 1, \ldots, n$ y $j = 1, \ldots, p$.

En las Figuras 6 y 7 se observa la estructura de correlación de los predictores.

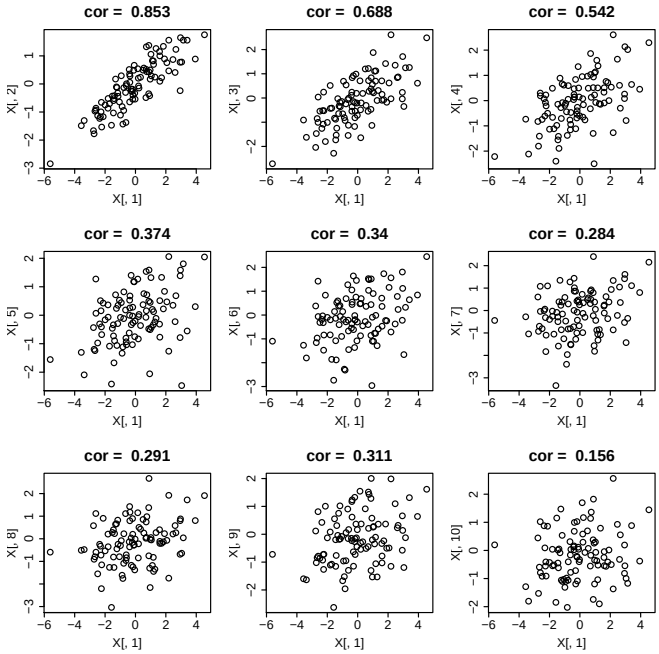

Figura 6: Diagrama de dispersión y coeficiente de correlación lineal entre $x_{\cdot 1}$ y $x_{\cdot j}$, para $j = 2, \ldots, 10$.

Luego se define el vector de coeficientes $\beta = (\beta_1, \ldots, \beta_p)$, donde $s = \#\{j : \beta_j \neq 0\} = 10$ y los valores para los predictores con efecto son: $\pm 1, \pm 2, \ldots, \pm 5$, cuyos índices son elegidos aleatoriamente (siendo los demás coeficientes nulos) (ver Figura 8).

Por último, se simula $\epsilon_i \sim N(0,1)$ independiente de x_{ij} y se obtienen n observaciones de la variable respuesta a través del modelo:

$$y_i = \sum_{j=1}^{p} \beta_j x_{ij} + \epsilon_i, \;\; i = 1, \ldots, n$$

El objetivo es estimar los parámetros β_j utilizando Ridge, LASSO y SCAD, a partir de la muestra de entrenamiento $\{(y_i, x_{i1}, \ldots, x_{ip}), i = 1, \ldots, n\}$. Observar que la estimación directa por mínimos cuadrados no es viable en este caso.

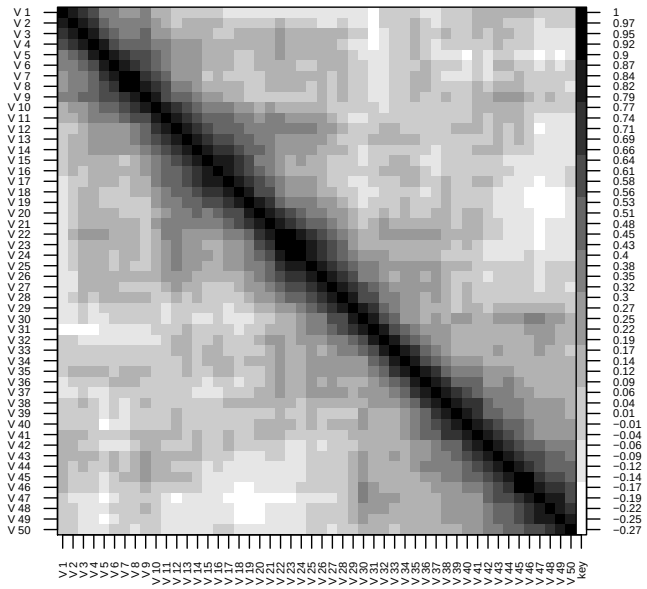

Figura 7: Coeficientes de correlación entre los 50 primeros predictores.

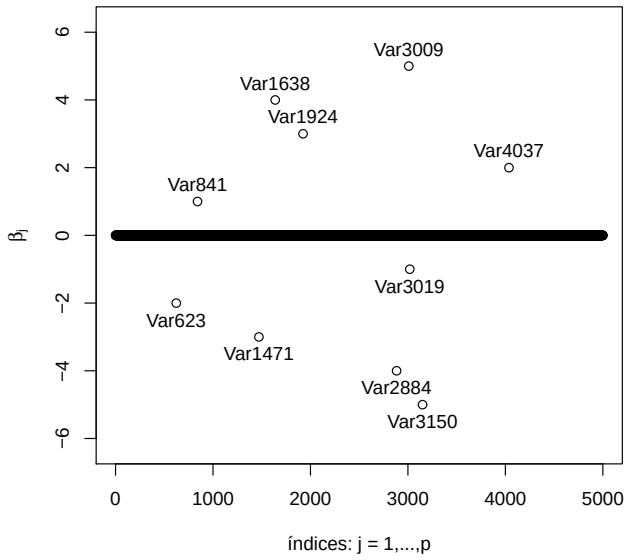

Figura 8: Coeficientes β_j de los predictores o variables.

5.1. Resultados

A continuación se presentan los resultados del ajuste mediante las técnicas Ridge, LASSO y SCAD. En los tres casos se comienza obteniendo el camino de soluciones

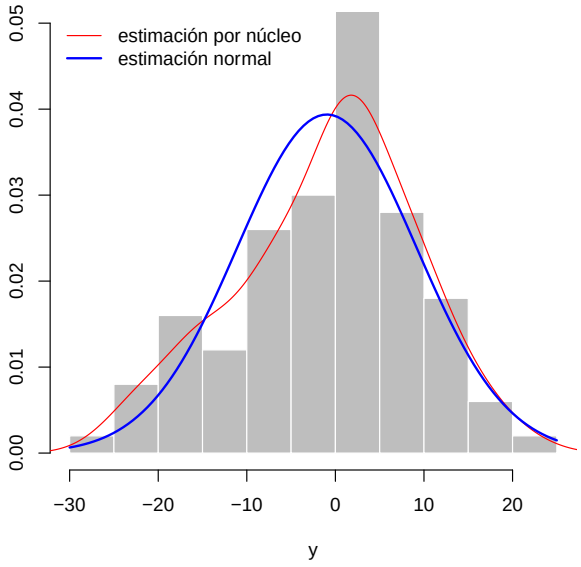

Figura 9: Histograma, estimación de densidad paramétrica (gaussiana) y no paramétrica (núcleo gaussiano) de la densidad de y.

$\left\{\hat{\beta}_j(\lambda) : \lambda \geq 0; j = 1, \ldots, p\right\}$ y luego se selecciona un modelo a través de *validación cruzada* (se obtiene el valor del parámetro de complejidad que minimiza una estimación del error de predicción).

5.1.1. Regresión Ridge

El camino de soluciones de Ridge muestra cómo las estimaciones se contraen hacia cero a medida que aumenta la penalización, pero sin anularse en ningún caso. Es decir, no se produce selección de variables. El modelo seleccionado mediante validación cruzada produce estimaciones muy pequeñas en relación a los coeficientes verdaderos (Figuras 12 y 13). La estructura esparsa del modelo verdadero no es capturada por este método.

5.1.2. Regresión LASSO

La estimación mediante LASSO anula algunos coeficientes produciéndose entonces selección de variables en forma automática. Sin embargo, para el modelo elegido mediante validación cruzada se incluyen 97 variables enre las cuales se encuentran 8 de las 10 con efecto real (las de mayor tamaño de efecto $|\beta_j|$). A su vez la contracción hacia cero de los coeficientes es menos pronunciada que en Ridge.

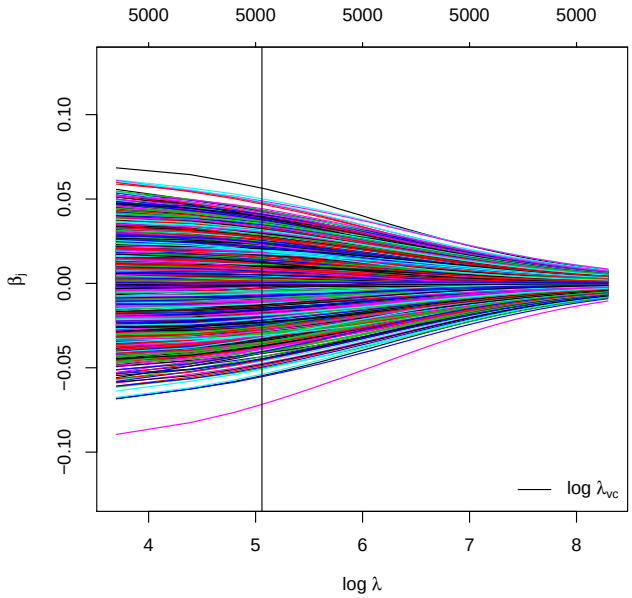

Figura 10: Camino de soluciones para Ridge.

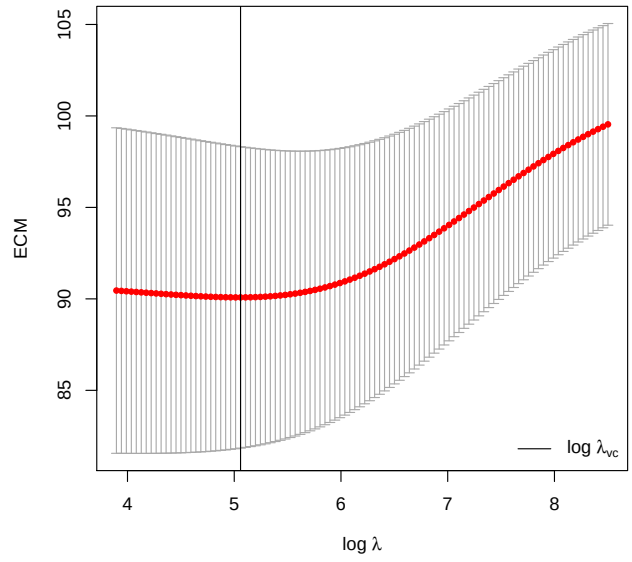

Figura 11: Estimación del error por validación cruzada para Ridge.

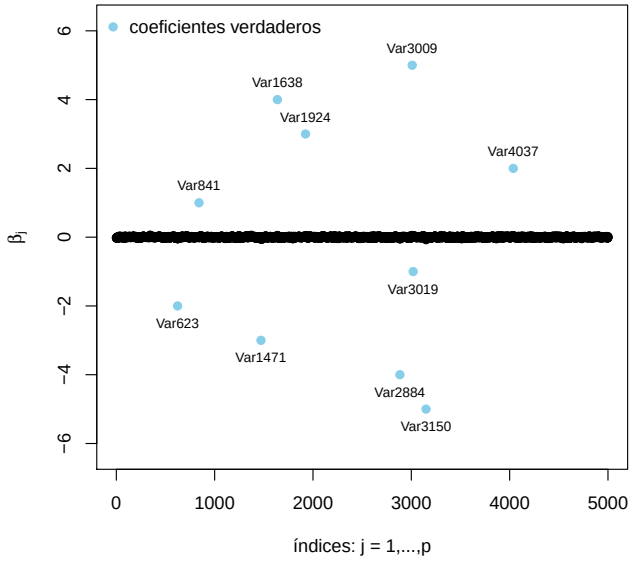

Figura 12: Coeficientes estimados por Ridge para modelo seleccionado por validación cruzada.

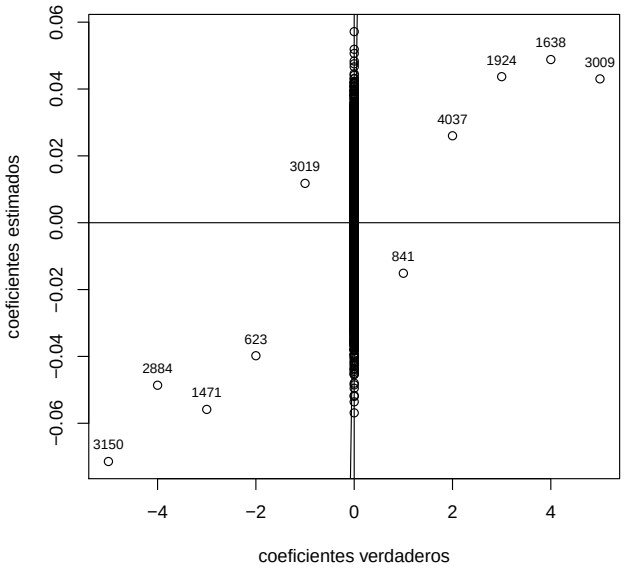

Figura 13: Comparación entre coeficientes verdaderos y estimados por Ridge.

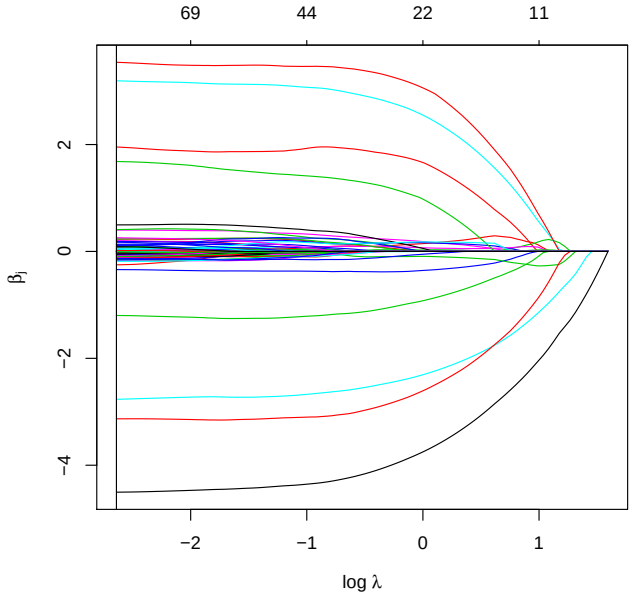

Figura 14: Camino de soluciones para LASSO.

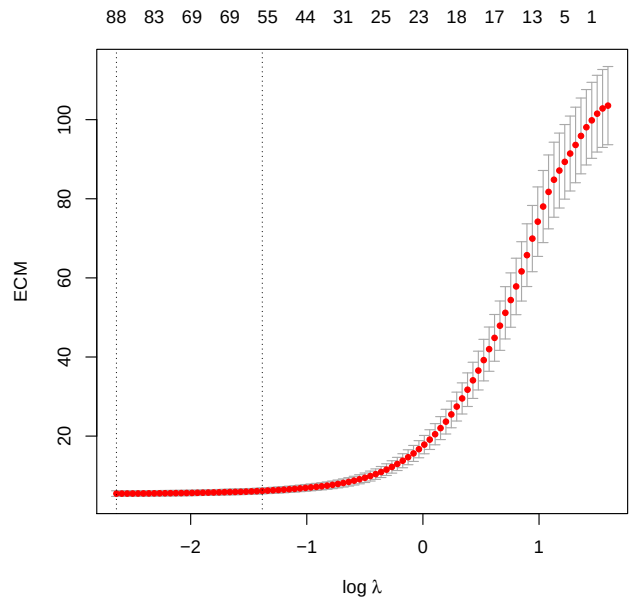

Figura 15: Estimación del error por validación cruzada para LASSO.

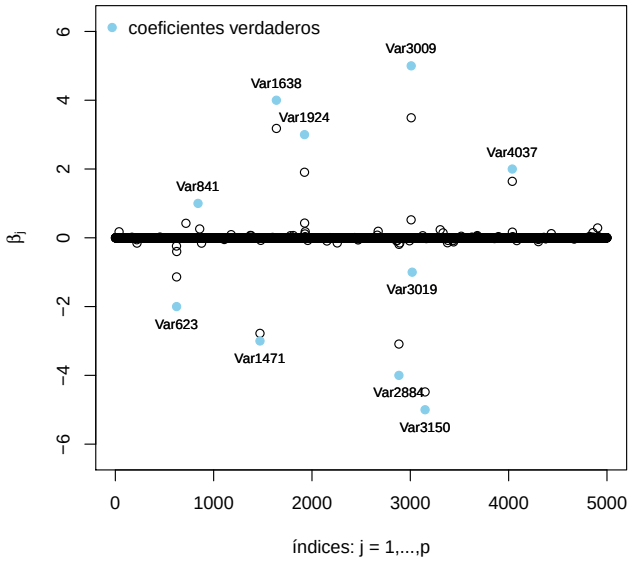

Figura 16: Coeficientes estimados por LASSO para modelo seleccionado por validación cruzada.

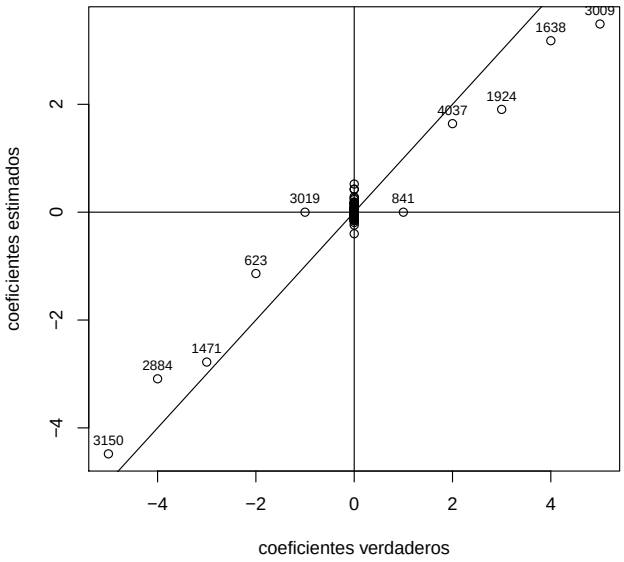

Figura 17: Comparación entre coeficientes verdaderos y estimados por LASSO.

5.1.3. SCAD

Por último, la utilización de la penalización SCAD produce un camino de soluciones más estable que en LASSO (menos variables son seleccionadas en el transcurso del mismo). El modelo seleccionado por validación cruzada incorpora 13 variables en total, 9 de las cuales tienen efecto real. A su vez, el efecto de contracción en la estimación es menos notorio debido a la propiedad de insesgadez mencionada anteriormente.

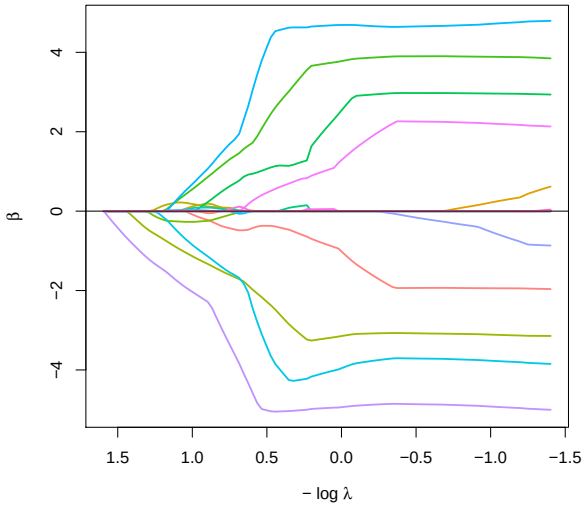

Figura 18: Camino de soluciones para SCAD.

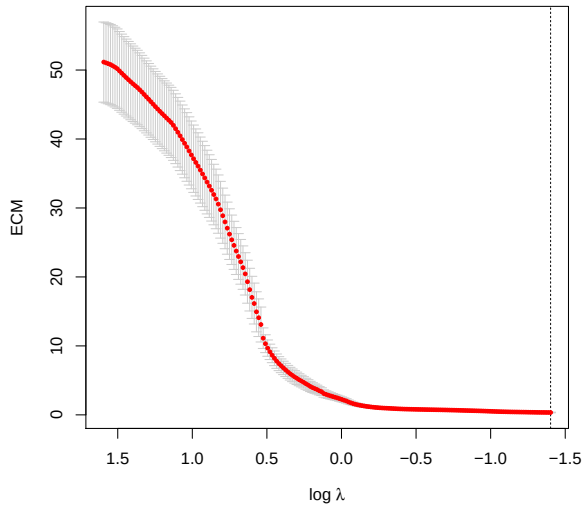

Figura 19: Estimación del error por validación cruzada para SCAD.

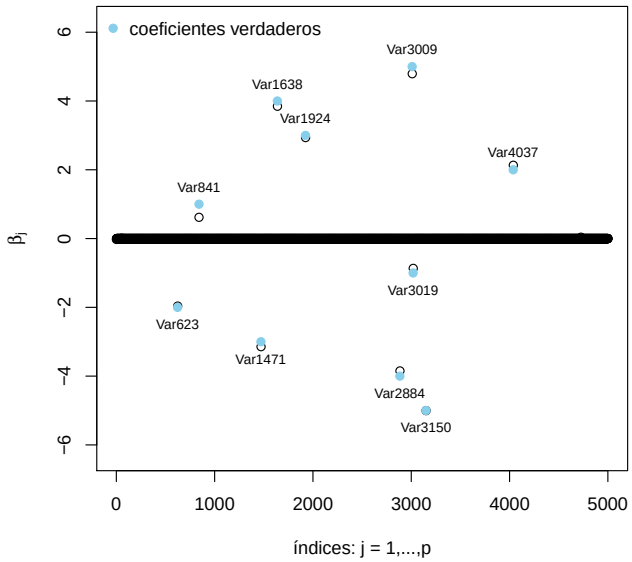

Figura 20: Coeficientes estimados por SCAD para modelo seleccionado por validación cruzada.

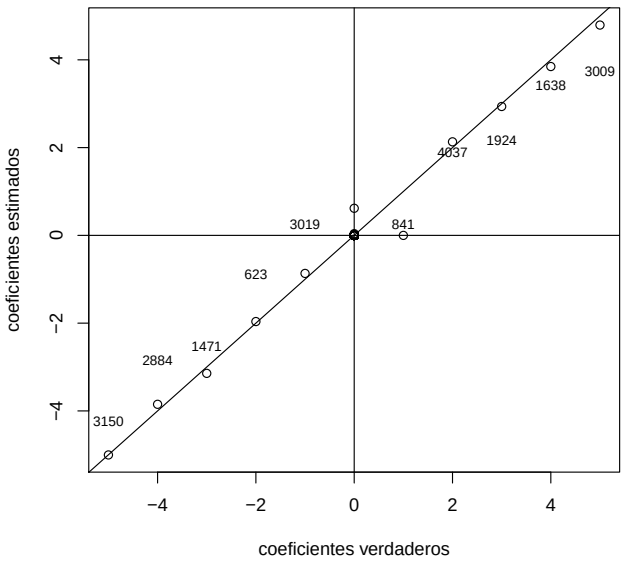

Figura 21: Comparación entre coeficientes verdaderos y estimados por SCAD.

5.1.4. Resumen de resultados

La estimación obtenida por Ridge, LASSO y SCAD se resume en el Cuadro 1 junto a los valores verdaderos del vector de coeficientes y el modelo *oráculo*, en el cual se estima por MCO utilizando únicamente las variables con efecto (este es el modelo de referencia en muchos casos para comparar los resultados de las diversas técnicas).

j	β_j	$\hat{\beta}_j^{orac}$	$\hat{\beta}_j^{Ridge}$	$\hat{\beta}_j^{LASSO}$	$\hat{\beta}_j^{SCAD}$
841	1	0.686	-0.015	0	0
4037	2	2.138	0.026	1.642	2.132
1924	3	2.850	0.044	1.906	2.937
1638	4	3.950	0.049	3.180	3.848
3009	5	4.836	0.043	3.490	4.793
3019	-1	-0.940	0.012	0	-0.866
623	-2	-1.941	-0.040	-1.137	-1.963
1471	-3	-3.092	-0.056	-2.779	-3.144
2884	-4	-3.866	-0.049	-3.089	-3.849
3150	-5	-5.018	-0.071	-4.482	-5.003
Nºvars.	10	10	5000	97	13

Cuadro 1: Comparación entre coeficientes verdaderos y estimados para el modelo *oráculo*, Ridge, LASSO y SCAD (se muestran solo los correspondientes a variables con efecto real).

6. Comentarios finales y algunas posibles líneas de investigación a seguir

El estudio de técnicas de análisis de datos y, en particular de regresión, en grandes dimensiones es una de las áreas más dinámicas de investigación en los últimos años (Fan y Li, 2006; Johnstone y Titterington, 2009; Fan y Lv, 2010). En particular, el énfasis ha estado en el estudio a nivel teórico de las técnicas (análisis de consistencia y eficiencia asintótica de los estimadores, por ejemplo), el desarrollo de algoritmos computacionales eficientes y los distintos desafíos de la aplicación de las mismas en diversas áreas científicas. Permanecer al tanto de estos nuevos desarrollos y profundizar en sus diversos aspectos plantea desafíos para aquéllos interesados en estos temas.

En el área de Econometría en particular, además de la aplicación directa en modelos de regresión de este tipo de técnicas, Fan y Qi (2011) plantean la potencial aplicación *modelos de vectores autorregresivos* (VAR), *datos de panel* y estimación de *matrices de volatilidad* en finanzas. Por su parte, Belloni y Chernozhukov (2011) muestran una aplicación sobre modelos empíricos de crecimiento económico.

Referencias

[1] Belloni, A., Chernozhukov, V. (2011). *High Dimensional Sparse Econometric Models: An Introduction*, Inverse Problems and High-Dimensional Estimation, Lecture Notes in Statistics, Vol. 203, pp. 121-156.

[2] Belloni, A., Chernozhukov, V., Hansen, C. (2011). *Inference for High Dimensional Sparse Econometric Models*, Advances in Economics and Econometrics, 10th World Congress of Econometric Society.

[3] Breiman, L. (1996). *Heuristics of instability and stabilization in model selection*, The Annals of Statistics, Vol. 24, No. 6, 2350-2383.

[4] Celeux, G., El Anbari, M., Marin, J-M., Robert, C. (2012). *Regularization in Regression: Comparing Bayesian and Frequentist Methods in a Poorly Informtative Situation*, Bayesian Analysis, Vol. 7, No. 1, 1-26.

[5] Clarke, B, Fokoué, E., Zhang, H. (2009). *Principles and Theory for Data Mining and Machine Learning*, Springer.

[6] Donoho, D. (2000). *High-Dimensional Data Analysis: The Curses and Blessings of Dimensionality*, Lecture Notes at Math Challenges of the 21st Century.

[7] Efron, B., Hastie, T., Johnstone, I., Tibshirani, R. (2004). *Least Angle Regression*, Ann. Stat. Vol. 32, No. 2, 407-499.

[8] Fan, J., Li, R. (2001). *Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties*, Journal of the American Statistical Association, Vol. 96, No. 456, 1348-1360.

[9] Fan, J., Li, R. (2006). *Statistical challenges with high dimensionality: feature selection in knowledge discovery*, International Congress of Mathematicians, Madrid, España. Sociedad Matemática Europea.

[10] Fan, J., Lv, J. (2010). *A selective overview of variable selection in high dimensional feature space*, Statistica Sinica 20, 101-148.

[11] Fan, J., Lv, J., Qi, L. (2011). *Sparse high-dimensional models in economics*, Annual Review of Economics, 3, 291-317.

[12] Friedman, J., Hastie, T., Tibshirani, R. (2010). *Regularization Paths for Generalized Linear Models via Coordinate Descendent*, Journal of Statistical Software, Vol. 33, Issue 1, 1-22.

[13] Fu, W. (1998). *Penalized Regressions: The Bridge versus the Lasso*, Journal of Computational and Graphical Statistics, Vol. 7, No. 3, 397-416.

[14] George, E. (2000). *The Variable Selection Problem*, Journal of the American Statistical Association. Vol. 95, No. 452, 1304-1308.

[15] Hans, C. (2009). *Bayesian lasso Regression*, Biometrika. Vol. 96, Issue 4, 835-845.

[16] Hans, C. (2010). *Model uncertainty and variable selection in Bayesian lasso regression*, Stat. Comput. Vol. 20, 221-229.

[17] Hansen, B. (2005). *Challenges for Econometric Model Selection*, Econometric Theory, 21, 2005, 6068.

[18] Hastie,T., Tibshirani, R, Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. Springer, 2^{nd} Edition.

[19] Hoerl, A., Kennard, R. (1970). *Ridge Regression: Biased Estimation for Nonorthogonal Problems*, Technometrics.

[20] Izenman, A. (2008). *Modern Multivariate Statistical Techniques. Regression, Classification and Manifold Learning*, Springer.

[21] Johnstone, I., Titterington, M. (2009). *Statistical challenges of high-dimensional data*, The Royal Society, Philosophical Transactions.

[22] Kyung, M., Gill, J., Ghosh, M., Casella, G. (2010). *Penalized Regression, Standard Errors, and Bayesian Lassos*, Bayesian Analysis, Vol. 5, No. 2, 369-412.

[23] Li, X., Xu, R. eds. (2009). *High-Dimensional Data Analysis in Oncology*, Springer.

[24] Li, Q., Lin, N. (2010). *The Bayesian Elastic Net*, Bayesian Analysis, Vol. 5, No. 1, 151-170.

[25] Meinshausen, N. (2006). *Relaxed Lasso*, Computational Statistics and Data Analysis. Vol. 52, Issue 1, 374-393.

[26] Park, T., Casella, G. (2008). *The Bayesian Lasso*, J. A. Statist. Assoc. Vol. 103, No. 482, 681-686.

[27] R Development Core Team. 2012. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.r-project.org.

[28] Sheather, J. (2009). *A Modern Approach to Regression with R*, Springer.

[29] Tibshirani, R. (1996). *Regression Shrinkage and Selection via the Lasso*, J. R. Statist. Soc., Serie B., Vol. 58, No. 1, 267-288.

[30] Tibshirani, R. (2011). *Regression shrinkage and selection via the lasso: a retrospective*, J. R. Statist. Soc., Serie B, Vol. 73, Issue 3, 273-282.

[31] Varmuza, K., Filzmoser, P. (2009). *Introduction to Multivariate Statistical Analysis in Chemometrics*, CRC Press.

[32] Zou, H., Hastie, T. (2005). *Regularization and variable selection via the elastic net*, J. R. Statist. Soc., Serie B, Vol. 67, Part 2, 301-320.

[33] Zou, H. (2006). *The adaptive Lasso and its oracles properties*, J. Am. Statist., Vol. 101, 1418-1429.