

Estimación y selección de variables en grandes dimensiones

Regresión Ridge, GNN, Lasso, Elastic Net, SCAD ...

Sebastián Castro

Aprendizaje Automático y aplicaciones
IMERL / FING / Udelar

Lunes 10 de junio de 2013

Índice general

1 Introducción

- Breve repaso de regresión lineal.
- Estimación por Mínimos Cuadrados Ordinarios (MCO).
- Problemas con MCO.

2 Más allá de MCO

- Selección de variables.
- Estabilización mediante Ridge y Garrote No Negativo.

3 Técnicas de regularización

- Introducción: regresión Lasso
- Especificación, propiedades e implementación.
- Descripción: ajuste(datos) + λ^* complejidad(modelo).
- Extensiones a Modelos Lineales Generalizados.
- Una perspectiva bayesiana.

4 Software

5 Aplicación

6 Referencias

Sección actual

1 Introducción

- Breve repaso de regresión lineal.
- Estimación por Mínimos Cuadrados Ordinarios (MCO).
- Problemas con MCO.

2 Más allá de MCO

- Selección de variables.
- Estabilización mediante Ridge y Garrote No Negativo.

3 Técnicas de regularización

- Introducción: regresión Lasso
- Especificación, propiedades e implementación.
- Descripción: ajuste(datos) + λ^* complejidad(modelo).
- Extensiones a Modelos Lineales Generalizados.
- Una perspectiva bayesiana.

4 Software

5 Aplicación

6 Referencias

Breve repaso de regresión lineal

- Problema: se intenta describir y predecir el comportamiento de una variable $Y \in R$ a partir de un conjunto de variables $X = (X_1, \dots, X_p) \in R^p$, que se supone *a priori* podrían estar asociadas con Y .
- A su vez, habitualmente suele ser de interés poder indicar cuáles variables predictoras X_j se encuentran efectivamente asociadas con la respuesta Y , y si fuera posible, en qué forma la afectan.
- Asumiendo un modelo de error aditivo:

$$Y = f(X_1, \dots, X_p) + \epsilon \quad (1)$$

con ϵ independiente de X , $E(\epsilon) = 0$ y $Var(\epsilon) = \sigma^2$.

Breve repaso de regresión lineal

- Considerando una *función de pérdida cuadrática*, $L(Y, f(X)) = (Y - f(X))^2$, la función que minimiza el *riesgo* (pérdida esperada), $R(f) = E L(Y, f(X))$ para cada X , es la *función de regresión*: $f(X) = E(Y|X)$ (Hastie y otros, 2009).
- Una aproximación habitual consiste en considerar $f \in \mathcal{F}$, siendo $\mathcal{F}$ cierta clase de funciones especificadas de antemano, y dentro de ella encontrar la *mejor aproximación* de f en base a una muestra de entrenamiento (observaciones *iid* de tamaño n , proveniente de la distribución conjunta de (Y, X)).
- Uno de los enfoques más sencillos y habituales plantea especificar un *modelo de regresión lineal*:

$$Y_i = \sum_{j=1}^p \beta_j X_{ij} + \epsilon_i, \quad i = 1, \dots, n \quad (2)$$

Breve repaso de regresión lineal

- Este modelo presenta ventajas desde el punto de vista de la estimación y la interpretación, al tiempo que es posible incorporar en el mismo *efectos no lineales* en las variables (X_j^2 , $\log X_j$, $\sqrt{X_j}$) y *términos de interacción* ($X_j X_k$) si los mismos son sospechados de antemano (a costo de incrementar el número de parámetros en el modelo).
- Por otro lado, si la relación entre la variable respuesta y los predictores es suficientemente compleja y desconocida, el modelo lineal puede no ser adecuado y otros modelos más flexibles deberían ser considerados (Breiman, 2001).

Estimación por Mínimos Cuadrados Ordinarios (MCO)

- A partir de una muestra de entrenamiento,
 $\{(y_i, x_{ij}) : i = 1, \dots, n; j = 1, \dots, p\}$, se plantea el modelo (2)
en términos matriciales: $Y = X\beta + \epsilon$, donde
 $Y = (y_1, \dots, y_n)^{tr}$, $X = ((x_{ij}))$, $\beta = (\beta_1, \dots, \beta_p)^{tr}$ y
 $\epsilon = (\epsilon_1, \dots, \epsilon_n)^{tr}$.
- La estimación habitual por *mínimos cuadrados* de β se realiza
minimizando la suma de cuadrados de los residuos:

$$SCR(\beta) = \|Y - X\beta\|_2^2 = \sum_{i=1}^n \left(y_i - \sum_{j=1}^p \beta_j x_{ij} \right)^2 \quad (3)$$

Algunas propiedades del estimador MCO

- Derivando $SCR(\beta)$ e igualando a 0 se obtiene el *sistema de ecuaciones normales*: $X^{tr}X\beta = X^{tr}Y$. Como la función objetivo es *convexa* y *diferenciable*, de esta manera se obtiene un mínimo.
- En el caso de que las columnas de X formen un conjunto linealmente independiente, la solución es única y está dada por:

$$\hat{\beta}^{mco} = (X^{tr}X)^{-1}X^{tr}Y \quad (4)$$

- Bajo los supuestos de X fija, $E(\epsilon) = 0$, $Var(\epsilon) = \sigma^2 I_n$, puede mostrarse que $\hat{\beta}^{mco}$ es el estimador de mínima variancia de β en la clase de estimadores lineales e insesgados, estimador BLUE por sus siglas en inglés (Teorema de Gauss-Markov).

Algunas propiedades del estimador MCO

- Con los supuestos adicionales de normalidad, $\epsilon \sim N(0, \sigma^2 I_n)$, se obtiene que $\hat{\beta}^{mco} \sim N(\beta, \sigma^2 (X^{tr} X)^{-1})$, con lo cual es posible hacer inferencia (prueba de hipótesis e intervalos de confianza por ejemplo) para β o funciones de β .
- Adicionalmente, bajo estos supuestos puede verificarse fácilmente que la *Estimación Máximo Verosímil* (EMV) de β coincide con la de mínimos cuadrados.

Algunos problemas con MCO

- Existen al menos dos razones por las cuales el estimador $\hat{\beta}^{mco}$ podría no ser adecuado en ciertas situaciones (Tibshirani, 1996):
 - (a) *baja precisión en las predicciones*; el estimador a menudo presenta poco sesgo pero gran variancia, lo cual se traduce en un pobre poder predictivo sobre nuevas observaciones,
 - (b) *falta de interpretabilidad*; si se utiliza un gran número de predictores (necesario para tener bajo sesgo ante un problema más o menos complejo), sería deseable determinar un pequeño subconjunto de éstos con fuerte poder explicativo y predictivo.

Problemas con MCO

- Las dificultades asociadas con el primer punto se encuentran vinculadas al problema de invertir la matriz $X^{tr}X$. Las mismas son tanto del tipo numéricas (problemas de redondeo que se propagan), como estadísticas (inflación de variancia).
- A su vez, ambas desventajas del estimador por MCO están vinculadas a la existencia de predictores fuertemente correlacionados.

Problemas con MCO

- Adicionalmente, *el caso $p \gg n$* (muchas más variables que observaciones) agrava estas dificultades ya que en ese caso el estimador no está bien definido (el sistema de ecuaciones normales es indeterminado).
- Esta situación es cada vez más frecuente en diversos ámbitos de la ciencia (Genética, Bioinformática, Procesamiento de Señales, Econometría, etc), con lo cual se ha convertido en un área de investigación muy dinámica en los últimos años (Fan y Li, 2006, Hastie y otros, 2009, Li y Xu, 2009).

En Uruguay no hemos estado al margen de estos temas...

- Curso: *Sparsity and Model Selection*, Jean-Marc Azaïs, Yohann de Castro, Fabrice Gamboa y Guillaume Obozinski. 28 de febrero al 4 de marzo de 2011, Centro de Matemática (CMAT), Facultad de Ciencias, UdelaR:
<http://www.math.univ-toulouse.fr/~decastro/Curso/SummerSchool.html>
- Curso: *Exploiting sparsity in high-dimensional statistical inference*, Arnak Dalalyan. 26 al 30 de noviembre de 2012, Escuela CIMPA *New trends in Mathematical Statistics*, Punta del Este:
<http://www.cmat.edu.uy/cmat/eventos/cimpa-stats>
- Curso: *Métodos Estadísticos para Predicción Genómica*, Gustavo de los Campos, Daniel Gianola y Santiago Avendaño. 20 y 21 de diciembre de 2012, Facultad de Agronomía, UdelaR.

Sección actual

1 Introducción

- Breve repaso de regresión lineal.
- Estimación por Mínimos Cuadrados Ordinarios (MCO).
- Problemas con MCO.

2 Más allá de MCO

- Selección de variables.
- Estabilización mediante Ridge y Garrote No Negativo.

3 Técnicas de regularización

- Introducción: regresión Lasso
- Especificación, propiedades e implementación.
- Descripción: ajuste(datos) + λ^* complejidad(modelo).
- Extensiones a Modelos Lineales Generalizados.
- Una perspectiva bayesiana.

4 Software

5 Aplicación

6 Referencias

Selección de variables

- Una solución habitual implica hacer *selección de variables* para obtener un modelo más *parsimonioso* y *estable* (George, 2000).
- Una primera aproximación consiste en ajustar los 2^p modelos posibles y comparar los mejores de cada tamaño $k \in \{1, \dots, p\}$.
- La comparación se realiza a través de alguna medida que tome en cuenta el *ajuste* a los datos de entrenamiento pero que penalice por la *complejidad* del modelo de forma tal que posea buen poder predictivo (generalización sobre datos nuevos).

Selección de variables

- Algunos ejemplos clásicos son:

$$R_{ajust}^2 = 1 - \frac{SCR(\beta(k))/(n - k - 1)}{SCT/(n - 1)}$$

$$AIC = n \log \left(\frac{SCR(\beta(k))}{n} \right) + 2k$$

$$BIC = n \log \left(\frac{SCR(\beta(k))}{n} \right) + (\log n)k$$

donde $SCT = \sum_{i=1}^n (y_i - \bar{y}_n)^2$ y k es el número de variables del modelo ajustado.

Selección de variables

- El método del mejor subconjunto de cada tamaño actualmente sólo es practicable si p no es demasiado grande ($p \approx 40$, $2^{40} \approx 1.0995 \times 10^{12}$) a través de algoritmos que utilizan la estructura anidada de los distintos modelos (*leaps and bound*).
- Cuando no es viable una búsqueda exhaustiva de todos los submodelos posibles, una opción razonable consiste en considerar un “buen camino” a través del espacio de modelos.
- Las técnicas más conocidas en estos casos son los métodos secuenciales o de a pasos (*stepwise*), en los cuales en el pasaje de un modelo a otro se agregan o eliminan variables de a una por vez (*Forward Selection*, *Backward Elimination* o *Forward-Backward*).

Selección de variables

- Estos son métodos *greedy* que reemplazan la búsqueda de un óptimo global por la consideración sucesiva de óptimos locales, con lo cual no garantizan la mejor solución y ni siquiera la misma entre sus distintas variantes.
- Sin embargo, la mayor desventaja que poseen es su fuerte *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).
- Esto se debe principalmente a que realizan un *proceso discreto* de exploración del espacio de modelos (cada variable es seleccionada o descartada).

Regresión Ridge

- Esta técnica fue propuesta originalmente en los años setenta, como un método para lidiar con el problema de colinealidad en un modelo lineal estimado por mínimos cuadrados, aún en el contexto $p < n$ (Hoerl y Kennard, 1970).
- Recordando que $\hat{\beta}^{mco} = (X^{tr}X)^{-1}X^{tr}y$ 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 $k > 0$ a cada término de la diagonal de $X^{tr}X$ antes de invertir la matriz.

Regresión Ridge

- Este proceso resulta en el *estimador ridge*:

$$\hat{\beta}^{ridge}(k) = (X^tr X + kI_p)^{-1} X^tr y \quad (5)$$

- El principal problema a resolver en la aplicación de Regresión Ridge es la determinación del valor de k más adecuado. La elección de este parámetro 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 del estimador, y la determinación final implica un compromiso entre ambos términos (Izenman, 2008).

Regresión Ridge

- Un método inicial y que aún continúa siendo sugerido por diversos autores, consiste en graficar simultáneamente los coeficientes de regresión estimados en función de k , y elegir el menor valor del parámetro para el cual se estabilizan dichos coeficientes.
- Un método más automático, pero intensivo computacionalmente, consiste en estimar k mediante *validación cruzada*. En general se recomienda utilizar ambos métodos y comparar los resultados.

Regresión Ridge

- Una derivación alternativa del estimador Ridge está dada por el siguiente problema de optimización con restricciones:

$$\min_{\beta} \left\{ \sum_{i=1}^n \left(y_i - \sum_{j=1}^p \beta_j x_{ij} \right)^2 \right\}, \text{ sujeto a } \sum_{j=1}^p \beta_j^2 \leq s \quad (6)$$

- El cual puede escribirse también en su versión *Lagrangiana*:

$$\hat{\beta}^{ridge} = \operatorname{argmin}_{\beta} \left\{ \sum_{i=1}^n \left(y_i - \sum_{j=1}^p \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^p \beta_j^2 \right\} \quad (7)$$

siendo $s, \lambda \geq 0$ los respectivos parámetros de penalización por complejidad.

Regresión Ridge

- Para evitar que la penalización varíe frente a cambios de escala de las variables, habitualmente éstas son estandarizadas (media 0 y variancia 1), aunque algunos autores prefieren analizar en cada caso si es lo más adecuado (Izenman, 2008).
- Las expresiones (6) y (7) muestran que el estimador de Ridge realiza un balance entre sesgo y variancia controlando el “tamaño” del vector de coeficientes mediante una penalización de norma L_2 . El estimador contrae los coeficientes β_j hacia cero respecto de los obtenidos por MCO (*shrinkage*).
- Observando que $SCR(\beta) = (y - X\beta)^{tr}(y - X\beta) = (\beta - \hat{\beta}^{mco})^{tr} X^{tr} X (\beta - \hat{\beta}^{mco}) + \text{constante}$, la expresión (6) puede visualizarse gráficamente en el caso de dos dimensiones.

Regresión Ridge

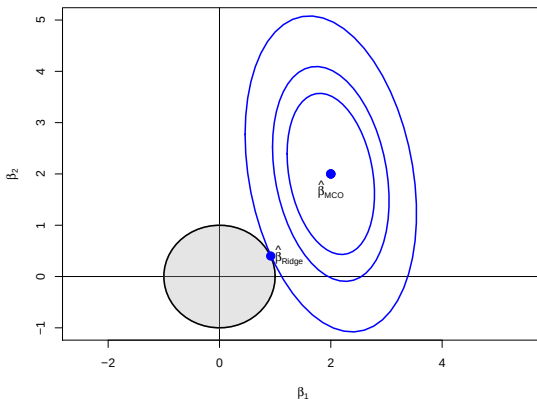


Figura : Descripción gráfica de la estimación Ridge en dos dimensiones.

Regresión Ridge

- En general, Regresión Ridge produce predicciones más precisas que los modelos obtenidos por MCO + selección “clásica” de variables, a menos que el verdadero modelo sea raro o “*esparsa*” (mayoría de coeficientes nulos).
- Sin embargo, si bien al aumentar λ (mayor penalización) los coeficientes estimados se contraen hacia cero, ninguno de ellos vale exactamente cero por lo cual no se produce selección de variables. Todas las variables originales permanecen en el modelo final.

Garrote No Negativo (GNN)

- Con el objetivo de encontrar un *compromiso* entre la simplicidad de obtener un modelo a través de la selección de variables, y la estabilidad y la precisión de Regresión Ridge, Breiman (1995) propuso la técnica del *Garrote No Negativo*.
- La idea fue minimizar respecto de $c = (c_1, \dots, c_p)$, $s \geq 0$:

$$\sum_{i=1}^n \left(y_i - \sum_{j=1}^p c_j \hat{\beta}_j x_{ij} \right)^2 \quad \text{sujeito a } c_j \geq 0 \text{ y } \sum_{j=1}^p c_j \leq s \quad (8)$$

donde $\hat{\beta}_j$ son los estimadores obtenidos por MCO.

Garrote No Negativo (GNN)

- Los valores de c_j son obtenidos resolviendo el problema de *programación cuadrática* (8).
- Luego, los coeficientes estimados por GNN son:

$$\hat{\beta}_j^{gnn} = c_j \hat{\beta}_j, \quad j = 1, \dots, p$$

- A medida que decrece s , la mayoría de los c_j se hacen cero y los restantes $\hat{\beta}_j^{gnn}$ no nulos son contraídos hacia cero.
- El parámetro de regularización s es determinado por validación cruzada con el propósito de minimizar el error de predicción esperado (Breiman, 1995).

Garrote No Negativo (GNN)

- La técnica del GNN elimina algunas variables, contrae otras y es relativamente estable (los resultados no cambian drásticamente con pequeñas modificaciones en los datos).
- Sin embargo, el estimador $\hat{\beta}^{gnn}$ depende de $\hat{\beta}^{mco}$ y, por lo tanto, no está bien definido cuando $p \gg n$ (situación no muy común por entonces, Tibshirani, 2011).
- Desarrollos más recientes permiten extender el uso de GNN en problemas de altas dimensiones, modificando el estimador inicial de MCO por uno más apropiado en este contexto (Yuan y Lin, 2007).

Sección actual

1 Introducción

- Breve repaso de regresión lineal.
- Estimación por Mínimos Cuadrados Ordinarios (MCO).
- Problemas con MCO.

2 Más allá de MCO

- Selección de variables.
- Estabilización mediante Ridge y Garrote No Negativo.

3 Técnicas de regularización

- Introducción: regresión Lasso
- Especificación, propiedades e implementación.
- Descripción: ajuste(datos) + λ^* complejidad(modelo).
- Extensiones a Modelos Lineales Generalizados.
- Una perspectiva bayesiana.

4 Software

5 Aplicación

6 Referencias

Regresión Lasso - Introducción

- También motivado por el objetivo de encontrar una técnica de regresión lineal que fuera estable pero que realizara selección de variables, Tibshirani (1996) propuso Lasso (Least Absolute Shrinkage and Selection Operator).
- Lasso es una técnica de regresión lineal *regularizada*, como Ridge, con una leve diferencia en la penalización (norma L_1 en lugar de L_2) que trae consecuencias importantes.
- El auge en los últimos años en la investigación y aplicación de técnicas tipo Lasso, se debe principalmente a la existencia de problemas donde $p \gg n$ y al desarrollo paralelo de algoritmos eficientes (Tibshirani, 2011).

Regresión Lasso - Especificación

- Lasso resuelve el problema de mínimos cuadrados con restricción sobre la norma- L_1 del vector de coeficientes:

$$\min_{\beta} \left\{ \sum_{i=1}^n \left(y_i - \sum_{j=1}^p \beta_j x_{ij} \right)^2 \right\}, \text{ sujeto a } \sum_{j=1}^p |\beta_j| \leq s \quad (9)$$

- O en forma equivalente, minimizando:

$$\sum_{i=1}^n \left(y_i - \sum_{j=1}^p \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^p |\beta_j| \quad (10)$$

siendo $s, \lambda \geq 0$ los respectivos parámetros de penalización por complejidad.

Regresión Lasso - Visualización

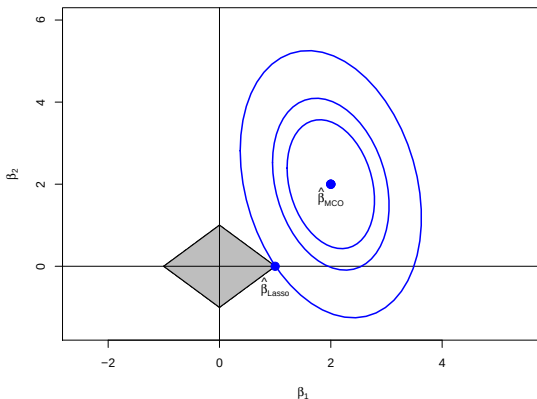


Figura : Descripción gráfica de la estimación Lasso en dos dimensiones.

Regresión Lasso - Algunas propiedades

- En forma similar a GNN y a diferencia de Ridge y MCO, el estimador de $\hat{\beta}^{lasso}$ es no lineal en el vector de respuesta Y , y no existe una expresión en forma “cerrada” del mismo (salvo en el caso de un diseño ortogonal $X^tr X = I_p$).
- Para valores crecientes de λ o decrecientes de s , los coeficientes β_j se contraen hacia cero como en Ridge (*shrinkage*), con la diferencia de que algunos de ellos se anulan.
- Esto es, Lasso produce estimación y selección de variables en forma continua y simultánea, siendo especialmente útil en el caso $p \gg n$.

Regresión Lasso - Implementación

- Los avances en los algoritmos para implementar Regresión Lasso en forma eficiente han sido muy importantes.
- En sus comienzos, la estimación se realizaba resolviendo para cada valor de s el problema de programación cuadrática (9). El método no es eficiente para un número grande de variables.
- Posteriormente, surgieron los algoritmos LARS (Efron y otros, 2004) y de *coordenada descendente* (Friedman y otros, 2010) que permitieron reducir enormemente el costo computacional (Tibshirani, 2011).

Técnicas de regularización

- Una formulación amplia de las técnicas de penalización/regularización puede plantearse como:

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \left\{ \sum_{i=1}^n \left(y_i - \sum_{j=1}^p \beta_j x_{ij} \right)^2 + \phi_{\lambda}(\beta) \right\} \quad (11)$$

donde $\beta = (\beta_1, \dots, \beta_p)$, $\lambda \geq 0$ y ϕ es una función de penalización sobre el “tamaño” de β , en general de la forma $\phi_{\lambda}(\beta) = \lambda \sum_{j=1}^p \phi_j(|\beta_j|)$ con ϕ_j creciente en $|\beta_j|$.

Técnicas de regularización

- Una familia de funciones de penalización muy utilizada es la correspondiente a la norma- L_q , dada por:

$$\phi_{\lambda}(\beta) = \lambda(\|\beta\|_q)^q = \lambda \sum_{j=1}^p |\beta_j|^q, q > 0$$

- Los estimadores resultantes en este caso son también conocidos como *estimadores Bridge* (Fu, 1998).

Técnicas de regularización

- Algunos casos particulares importantes son: $q = 1$ (Lasso) y $q = 2$ (Ridge). Además, los métodos que penalizan por el número de variables pueden ser vistos como el caso límite $q \rightarrow 0$.
- Para $q > 1$ el estimador no realiza selección de variables (Fan y Li, 2001). Por otro lado, Lasso corresponde al valor de q más pequeño que produce una región factible convexa.
- La convexidad del problema de optimización es deseable desde el punto de vista computacional. Funciones en varias variables no convexas pueden tener múltiples óptimos locales.

Técnicas de regularización

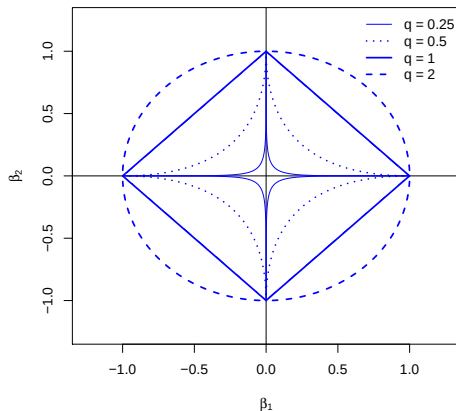


Figura : Curvas de nivel de la penalización L_q en dos dimensiones.

Otras técnicas de regularización

- En los últimos años se han presentado algunas generalizaciones y extensiones de las técnicas presentadas anteriormente, especialmente diseñadas para ciertas situaciones particulares.
- 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.
- A continuación se presentan algunas de las más importantes, acompañadas de una breve descripción y referencias respectivas.

Regresión Elastic Net

- Zou y Hastie, 2005, propusieron *Elastic Net* como un método de penalización que combina los beneficios de Ridge y Lasso.
- En primer lugar, se define el estimador *ingenuo* de Elastic Net, $\hat{\beta}$, como el que minimiza:

$$\sum_{i=1}^n \left(y_i - \sum_{j=1}^p \beta_j x_{ij} \right)^2 + \lambda_1 \sum_{j=1}^p |\beta_j| + \lambda_2 \sum_{j=1}^p \beta_j^2 \quad (12)$$

donde $\lambda_1, \lambda_2 \geq 0$ son ambos parámetros de complejidad.

Regresión Elastic Net

- Debido a que la doble penalización en (12) puede introducir sesgo en la estimación, se corrige el estimador anterior obteniéndose $\hat{\beta}^{enet} = (1 + \lambda_2)\hat{\beta}$ (Zou y Hastie, 2005).
- En cierta forma, el estimador de Elastic Net combina las fortalezas de Lasso (la penalización L_1 promueve soluciones esparsas), y de Ridge (predictores altamente correlacionados presentan coeficientes estimados similares).
- Existen algoritmos eficientes del tipo LARS (LARS-EN) y de coordenada descendente, para su implementación (Zou y Hastie, 2009, Friedman y otros, 2010).

Regresión Elastic Net

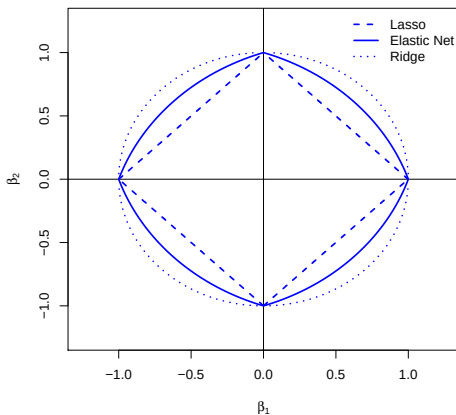


Figura : Curvas de nivel de penalización Lasso, Ridge y Elastic Net.

Regresión Lasso Adaptativo (ALasso)

- Esta técnica se encuentra motivada en el hecho de que bajo ciertas condiciones el estimador de Lasso no es consistente como método de selección de variables (Zou, 2006).
- Lasso Adaptativo (ALasso) es una generalización de Lasso que permite aplicar diferentes penalizaciones a las variables mediante la asignación de pesos distintos, los cuales dependen de los datos.
- Esta generalización permite imponer mayores penalizaciones sobre variables poco importantes y pequeñas penalizaciones sobre las más relevantes.

Regresión Lasso Adaptativo (ALasso)

- En ALasso, el problema consiste en minimizar respecto de β la expresión:

$$\sum_{i=1}^n \left(y_i - \sum_{j=1}^p \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^p w_j |\beta_j| \quad (13)$$

donde $w_j = 1/|\tilde{\beta}_j|^\gamma$, $j = 1, \dots, p$, son pesos positivos que aseguran propiedades de consistencia del estimador ALasso, $\gamma > 0$ es un parámetro de ajuste adicional y $\tilde{\beta}_j$ es un estimador inicial de β_j , por ejemplo mediante MCO, Ridge o el propio Lasso (Zou, 2006).

Regresión Lasso Relajado (Relaxed Lasso)

- Es un procedimiento en dos etapas, propuesto como una generalización de Lasso y especialmente diseñado para problemas de regresión en altas dimensiones (Meinshausen, 2006).
- En una primera etapa y para $\lambda > 0$ fijo, se aplica Regresión Lasso sobre el modelo completo y se define:

$$S(\lambda) = \left\{ j : \hat{\beta}(\lambda)^{lasso} \neq 0 \right\}$$

Regresión Lasso Relajado (Relaxed Lasso)

- Luego, el estimador de Lasso Relajado, $\hat{\beta}_j(\lambda, \phi)$, se define para $\phi \in (0, 1]$ como:

$$\operatorname{argmin}_{\beta} \left\{ \left(\sum_{i=1}^n y_i - \sum_{j \in S(\lambda)} \beta_j x_{ij} \right)^2 + \phi \lambda \|\beta_{S(\lambda)}\|_1 \right\} \quad (14)$$

- El parámetro λ regula la parte de selección de variables como en Lasso, mientras que el parámetro de *relajación* ϕ controla la contracción de los coeficientes.

Regresión Lasso Relajado (Relaxed Lasso)

- Si $\phi = 1$ la estimación coincide con Lasso mientras que si $\phi < 1$ la contracción de los coeficientes en el modelo seleccionado es menor que en Lasso.
- Los parámetros λ y ϕ pueden ser elegidos por validación cruzada.
- La estimación de los coeficientes se puede obtener en forma eficiente a través de un algoritmo basado en LARS (Meinshausen, 2006).

Penalizaciones no convexas (SCAD)

- Fan y Li (2001) proponen 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 .

Penalizaciones no convexas (SCAD)

- 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.

Penalizaciones no convexas (SCAD)

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

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

donde $a > 2$ y $\lambda > 0$ son parámetros de ajuste.

- El estimador de SCAD, $\hat{\beta}^{scad}$, se define entonces como el que minimiza:

$$\sum_{i=1}^n \left(y_i - \sum_{j=1}^p \beta_j x_{ij} \right)^2 + \sum_{j=1}^p \phi_{\lambda}(\beta_j) \quad (15)$$

Penalizaciones no convexas (SCAD)

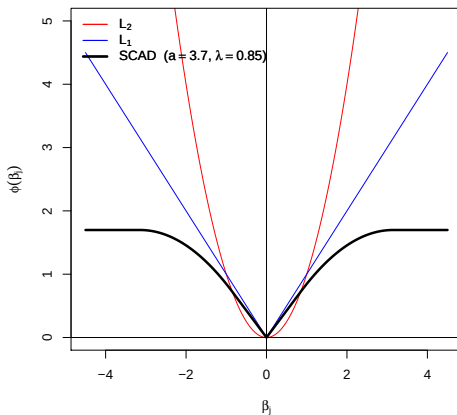


Figura : Funciones de penalización de Ridge (L_2), Lasso (L_1) y SCAD.

Penalizaciones no convexas (SCAD)

- 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 ilustra la diferencia entre ambas en la propiedad de insesgadez.
- 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 locales de la función objetivo (Fan y Li, 2001, Clarke y otros, 2009).

Extensiones a Modelos Lineales Generalizados (GLM)

- Las técnicas de penalización en regresión pueden extenderse a una amplia variedad de tipos de variable respuesta, incluyendo respuestas binarias, de conteo y continuas.
- 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).

Extensiones a Modelos Lineales Generalizados (GLM)

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

$$\sum_{i=1}^n \log f_i((y_i|g(\mathbf{x}_i^{tr}\beta))) - n \sum_{j=1}^p \phi_\lambda(\beta_j) \quad (16)$$

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

$$- \sum_{i=1}^n \log f_i((y_i|g(\mathbf{x}_i^{tr}\beta))) + n \sum_{j=1}^p \phi_\lambda(\beta_j) \quad (17)$$

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, si $y|\beta, \sigma^2, X \sim N(X\beta, \sigma^2 I_n)$, $p(\beta, \sigma^2|X) \propto \sigma^{-2}$, entonces:

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

Una perspectiva bayesiana sobre las técnicas de regularización

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

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

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

$$\sigma^2|y, X \sim Inv - \chi^2(n - p, s^2)$$

$$\text{con } s^2 = (y - X\hat{\beta}^{mco})^{tr}(y - X\hat{\beta}^{mco})/(n - p)$$

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

Una perspectiva bayesiana sobre las técnicas de regularización

- Utilizando distintas distribuciones a priori informativas, varias de las técnicas de regularización presentadas pueden ser vistas como estimadores bayesianos.
- Ridge: si a priori $\beta|\sigma^2 \sim N(0, \sigma^2 I_p)$, con $\lambda = \sigma^2/\sigma_\beta^2$, entonces:

$$\beta|y, \sigma^2, X \sim N(\hat{\beta}^{ridge}, \sigma^2(X^{tr}X + \lambda I_p)^{-1})$$

- Lasso: si $p(\beta|\lambda) = \prod_{j=1}^p p(\beta_j|\lambda)$, con

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

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

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

Una perspectiva bayesiana sobre las técnicas de regularización

- Con lo cual $\hat{\beta}^{lasso}$ coincide con el estimador *máximo a posteriori* (MAP) bajo este modelo.
- Otras técnicas de regularización pueden presentarse de esta manera donde la penalización se corresponde con una distribución a priori adecuada.
- ¿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 “*esparcidad*” del modelo es en definitiva una a priori...

Una perspectiva bayesiana sobre las técnicas de regularización

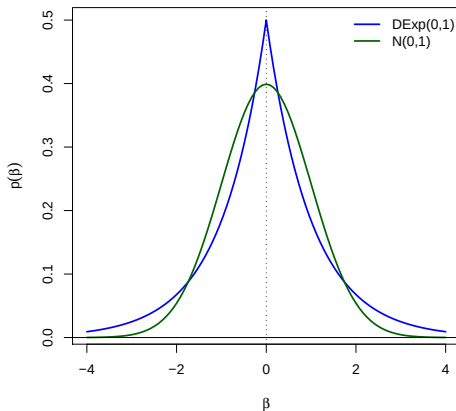


Figura : Ejemplos de distribuciones a priori implícitas en Ridge y Lasso.

Sección actual

1 Introducción

- Breve repaso de regresión lineal.
- Estimación por Mínimos Cuadrados Ordinarios (MCO).
- Problemas con MCO.

2 Más allá de MCO

- Selección de variables.
- Estabilización mediante Ridge y Garrote No Negativo.

3 Técnicas de regularización

- Introducción: regresión Lasso
- Especificación, propiedades e implementación.
- Descripción: ajuste(datos) + λ^* complejidad(modelo).
- Extensiones a Modelos Lineales Generalizados.
- Una perspectiva bayesiana.

4 Software

5 Aplicación

6 Referencias

Software

Técnica	Biblioteca
Ridge	<i>MASS, glmnet</i>
GNN	<i>lqa, oem</i>
Lasso	<i>lars, glmnet</i>
Elastic Net	<i>elasticnet, glmnet</i>
Alasso	<i>parcor</i>
Relaxo	<i>relaxo</i>
SCAD	<i>ncvreg</i>
Bayesian Lasso	<i>BLR</i>
Bayesian Ridge	<i>BLR</i>
Bayesian ENet	<i>BLR</i>

Cuadro : Bibliotecas disponibles en entorno R para el ajuste de algunas técnicas de regularización.

Sección actual

1 Introducción

- Breve repaso de regresión lineal.
- Estimación por Mínimos Cuadrados Ordinarios (MCO).
- Problemas con MCO.

2 Más allá de MCO

- Selección de variables.
- Estabilización mediante Ridge y Garrote No Negativo.

3 Técnicas de regularización

- Introducción: regresión Lasso
- Especificación, propiedades e implementación.
- Descripción: ajuste(datos) + λ^* complejidad(modelo).
- Extensiones a Modelos Lineales Generalizados.
- Una perspectiva bayesiana.

4 Software

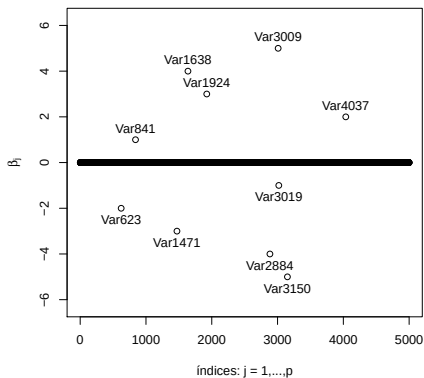
5 Aplicación

6 Referencias

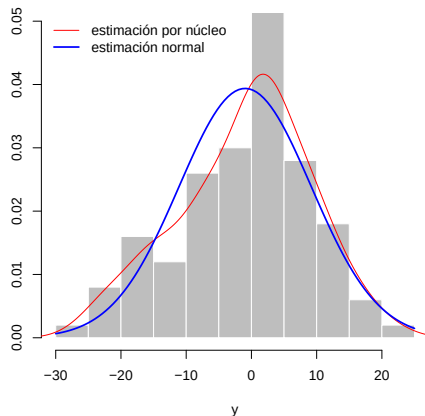
Simulación

- Se simulan $n = 100$ observaciones de un modelo lineal, $y = X\beta + \epsilon$, con $p = 5000$ variables ($p \gg n$).
- Predictores: $X = ((x_{ij}))$, donde $x_{ij} \sim N(0, 1)$, $\text{cor}(x_{.j}, x_{.k}) = \rho^{|j-k|}$ y $\rho = 0.85$, para $i = 1, \dots, n$ y $j = 1, \dots, p$.
- Coeficientes: $\beta = (\beta_1, \dots, \beta_p)$, donde $s = \#\{j : \beta_j \neq 0\} = 10$, con valores para predictores con efecto $\pm 1, \pm 2, \dots, \pm 5$, cuyos índices son elegidos aleatoriamente (demás coeficientes nulos).
- Por último, $\epsilon_i \sim N(0, 1)$ independiente de x_{ij} .

Coeficientes β_j de los predictores



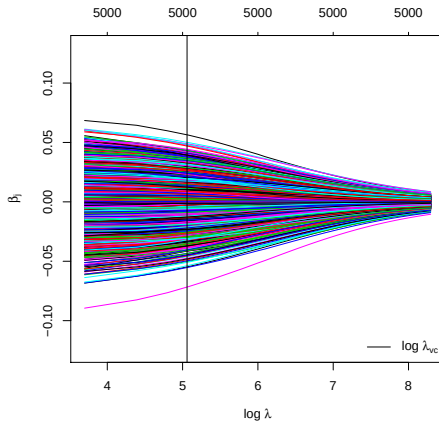
Distribución observada de variable respuesta



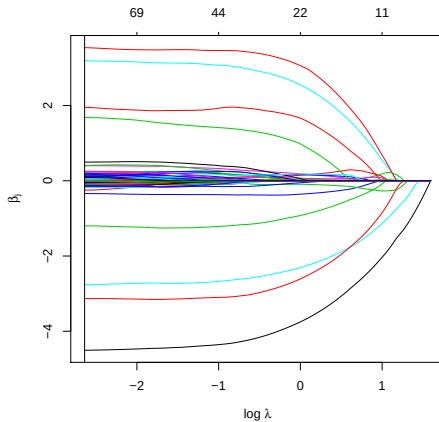
Ajuste de técnicas de regularización

- El objetivo es estimar los parámetros β_j utilizando Ridge, LASSO y SCAD, a partir de la muestra de entrenamiento $\{(y_i, x_{i1}, \dots, x_{ip}), i = 1, \dots, n\}$.
- En los tres casos se comienza obteniendo el camino de soluciones $\{\hat{\beta}_j(\lambda) : \lambda \geq 0; j = 1, \dots, p\}$ y luego se selecciona un modelo a través de *validación cruzada*.
- Observar que la estimación directa por mínimos cuadrados no es viable en este caso.

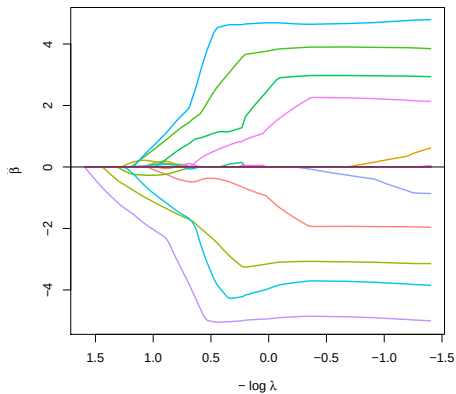
Camino de soluciones para Ridge



Camino de soluciones para LASSO



Camino de soluciones para SCAD



Comparación entre modelo oráculo, Ridge, LASSO y SCAD

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
Nº vars.	10	10	5000	97	13

Algunos posibles temas de interés a estudiar

- Aspectos matemáticos y estadístico-matemáticos: consistencia de estimadores, inferencia.
- Aspectos computacionales: algoritmos (LARS, coordenada descendente).
- Otras técnicas no mencionadas: Grouped and Fused Lasso, Dantzig Selector ...
- Conexiones de Lasso con Boosting.
- Análisis bayesiano (en especial, técnicas MCMC para la implementación).

Sección actual

1 Introducción

- Breve repaso de regresión lineal.
- Estimación por Mínimos Cuadrados Ordinarios (MCO).
- Problemas con MCO.

2 Más allá de MCO

- Selección de variables.
- Estabilización mediante Ridge y Garrote No Negativo.

3 Técnicas de regularización

- Introducción: regresión Lasso
- Especificación, propiedades e implementación.
- Descripción: ajuste(datos) + λ^* complejidad(modelo).
- Extensiones a Modelos Lineales Generalizados.
- Una perspectiva bayesiana.

4 Software

5 Aplicación

6 Referencias

Referencias

-  Breiman, L., (1995) *Better Subset Regression Using the Nonnegative Garrote*, Technometrics.
-  Breiman, L., (1996) *Heuristics of instability and stabilization in model selection*, The Annals of Statistics.
-  Breiman, L., (2001) *Statistical Modeling: The Two Cultures*, Statistical Science.
-  Clarke, B, Fokoué, E., Zhang, H. (2009) *Principles and Theory for Data Mining and Machine Learning*, Springer.
-  Efron, B., Hastie, T., Johnstone, I., Tibshirani, R. (2004) *Least Angle Regression*, Ann. Stat.

Referencias

-  Fan, J., y Li, R. (2001) *Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties*, Journal of the American Statistical Association.
-  Fan, J. y Li, R., (2006) *Statistical challenges with high dimensionality: feature selection in knowledge discovery*, International Congress of Mathematicians, Madrid, España. Sociedad Matemática Europea.
-  Friedman, J., Hastie, T., Tibshirani, R. (2010). *Regularization Paths for Generalized Linear Models via Coordinate Descent*, Journal of Statistical Software.
-  Fu, W., (1998) *Penalized Regressions: The Bridge Versus the Lasso*, Journal of Computational and Graphical Statistics.

Referencias

-  George, E. (2000) *The Variable Selection Problem*, Journal of the American Statistical Association.
-  Hastie, T., Tibshirani, R., Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. Springer, 2nd Edition.
-  Hoerl, A., Kennard, R. (1970). *Ridge Regression: Biased Estimation for Nonorthogonal Problems*. Technometrics.
-  Izenman, A. (2008) *Modern Multivariate Statistical Techniques. Regression, Classification and Manifold Learning*, Springer.
-  Li, X., Xu, R. (2009) *High-Dimensional Data Analysis in Oncology*, Springer.

Referencias



Meinshausen, N. (2006). *Relaxed Lasso*. Computational Statistics and Data Analysis.



R Development Core Team. 2011. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL <http://www.r-project.org>.



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



Tibshirani, R. (1996) *Regression Shrinkage and Selection via the Lasso*, J. R. Statist. Soc.

Referencias

-  Tibshirani, R. (2011) *Regression shrinkage and selection via the lasso: a retrospective*, J. R. Statist. Soc.
-  Varmuza, K., Filzmoser, P. (2009) *Introduction to Multivariate Statistical Analysis in Chemometrics*, CRC Press.
-  Yuan, M., Lin, Y. (2006) *On the non-negative garrotte estimator*, J. R. Statist. Soc. B.
-  Zou, H., Hastie, T. (2005) *Regularization and variable selection via the elastic net*, J. R. Statist. Soc.
-  Zou, H., (2006) *The adaptive Lasso and its oracles properties*, J. Am. Statist.