Shrinkage - 07/11

Metodi di Shrinkage

I metodi di Subset selection visti fino ad ora consistono nell'usare least squares per fare il fit di un modello lineare che contiene un sottoinsieme dei predittori.

In alternativa, possiamo fare il fit di un modello contenente tutti i p predittori usando una tecnica che *limita* (constrains) o *regolarizza* le stime dei coefficienti, o equivalentemente, che *restringe* (shrinks) le stime dei coefficienti verso zero.

Potrebbe non essere ovvio il perché una limitazione potrebbe migliorare il fit, ma dallo studio emerge che restringere (shrinking) le stime dei coefficienti può ridurre significativamente la loro varianza.

Le due tecniche più note per fare shrinking dei coefficienti della regressione verso zero sono: **ridge regression** (regressione a cresta) e **lasso** (lazo).

Ridge regression

Dato il training data-set $D = (x_i, y_i)_{i=1}^n$, OLS (ordinary least squares) stima $\beta_0, \ldots, \beta_p$ minimizzando:

$$\text{RSS} = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 .$$

La Ridge regression minimizza invece un'equazione leggermente diversa:

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

dove λ è un tuning parameter da determinare separatamente.

Il secondo termine, $\lambda \sum_{j=1}^{p} \beta_j^2$, è chiamato shrinkage penalty. Questo termine ha l'effetto di restringere (shrink) le stime di β_j verso 0.

Il parametro di tuning λ serve a controllare l'impatto relativo di questi due termini sulle stime dei coefficienti di regressione.

> 💡 La ridge regression produrrà un diverso set di stime di coefficienti, $\hat{\beta}_\lambda^R$, per ogni valore diverso di λ. Quando $\lambda = 0$, otteniamo proprio OLS.
>
> Risulta fondamentale selezionare un buon valore di λ.

Si nota che la shrinkage penalty (penalità di restringimento) è applicata a $\beta_1, \ldots, \beta_p$ ma non alla intercetta β_0.

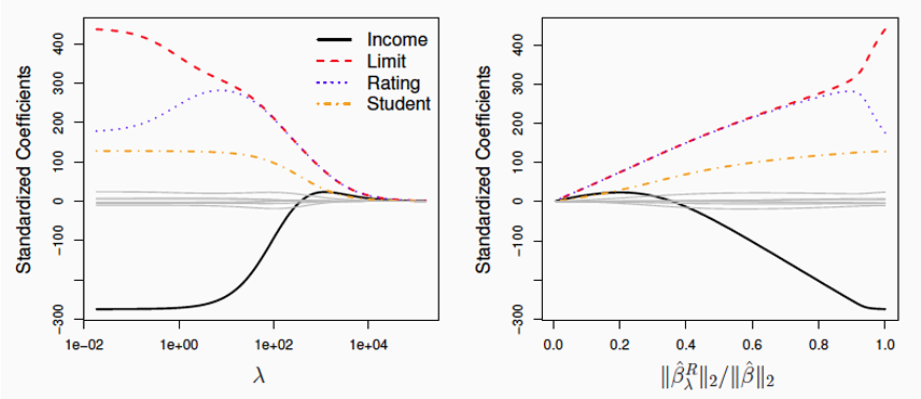
Noi non vogliamo restringere l'intercetta in quanto questa è semplicemente la misura della media della risposta quando le variabili indipendenti sono uguali a 0.

Per $\lambda \to \infty$ l'impatto del restringimento cresce e le stime dei coefficienti della ridge regression approcciano 0.

Credit data: ridge regression

Di seguito, sono mostrati i coefficienti standardizzati della ridge regression.

Al crescere di λ i coefficienti standardizzati si restringono verso 0.



Nel panel a sinistra ogni curva corrisponde alle stime dei coefficienti della ridge regression per una di dieci variabili, plotted in funzione di λ.

Nel panel a destra sono mostrate le stesse stime dei coefficienti della ridge regression ma invece di avere λ sull'asse x abbiamo $||\hat{\beta}^R_\lambda||_2 \ / \ ||\hat{\beta}||_2$ dove $\hat{\beta}$ denota il vettore delle stime dei coefficienti least squares.

La quantità $||\hat{\beta}^R_\lambda||_2 \ / \ ||\hat{\beta}||_2$ è nel range tra

- 1, quando $\lambda = 0$, i coefficienti della ridge regression sono gli stessi di Least Squares, e
- 0, quando $\lambda = \infty$, le stime dei coefficienti della ridge regression sono un vettore di 0.

Si può considerare questa quantità come la misura di quanto i coefficienti della ridge regression sono stati shrunken verso 0, un valore piccolo indica che sono stati ristretti quasi vicino a 0.

Ridge regresson: scaling (ridimensionamento) dei predittori

La stime OLS sono scale equivariant (equivariante alla scala): moltiplicando X_j per una costante c porta ad uno scaling delle stime dei coefficienti di un fattore $1/c$, cioè, $X_j\hat{\beta}_j$ resterà lo stesso.

Le stime di ridge regression possono cambiare sostanzialmente quando si moltiplica un dato predittore per una costante a causa del termine di shrinkage penalty. Specificatamente, $X_j\hat{\beta}^R_{j,\lambda}$, dipenderà non solo dal valore di λ ma anche dallo scaling dello j-esimo predittore e addirittura anche dallo scaling degli altri predittori.

Di conseguenza è meglio standardizzare i predittori prima di applicare la ridge regression:

$$\tilde{x}_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j}$$

$$\text{where } \bar{x}_j = \sum_{i=1}^{n} x_{ij}/n \text{ and } s_j^2 = \sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2/(n-1).$$

Allora la intercetta stimata è $\hat{\beta}_0 = \overline{y}$.

Perché la ridge regressione è migliore di OLS?

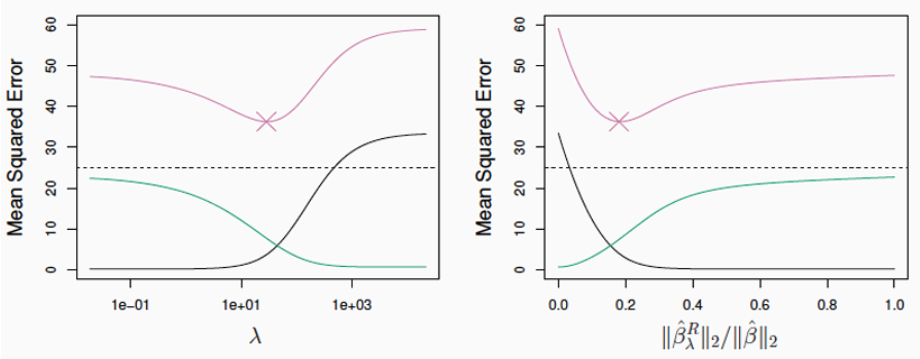
> 💡 C'è un bias-variance trade-off: il termine di penalità rende più biased le stime di ridge regression all'aumentare di λ ma ne può anche ridurre sostanzialmente la varianza.

In generale le stime ridge regression avranno più bias di OLS ma meno varianza.

Ne consegue che ridge regression funzionerà al meglio in situazioni nelle quali le stime OLS hanno alta varianza. Tra queste situazioni ricordiamo $p \cong n$ e la multicollinearity.

Qui di seguito è illustrato il concetto usando un data set simulato contenente p=45 predittori e n=50 osservazioni.

Linea nera: bias al quadrato. Linea verde: varianza. Linea viola: test MSE. Linea tratteggiata indica il minimo MSE possibile. Le crocette sulle linee viola indicano i modelli di ridge regression per i quali l'MSE è il più basso possibile.



La curva verde nel panel a sinistra mostra la varianza delle predizioni di ridge regression in funzione di λ. Con le stime dei coefficienti di least squares, che corrisponde alla ridge regressione con $\lambda = 0$, la varianza è alta ma non c'è bias. Man mano che λ aumenta, lo shrinkage delle stime dei coefficienti

ridge porta ad una sostanziale riduzione nella varianza della predizione, con il prezzo di un leggere incremento in bias.

Ricordiamo che il test MSE è funzione della varianza più il bias al quadrato.

Per valori di λ fino a 10 la varianza decresce rapidamente con un aumento di bias molto piccolo. Di conseguenza l'MSE diminuisce considerevolmente al crescere di λ da 0 a 10. Oltre questo punto la decrescita della varianza causata dall'aumento di λ diminuisce e lo shrinkage dei coefficienti li porta ad essere significativamente sottostimati il che risulta in un grande aumento di bias.

Il minimo MSE è ottenuto all'incirca a $\lambda = 30$.

Risulta interessante notare che a causa dell'alta varianza, l'MSE associato con il fit least squares ($\lambda = 0$) è quasi tanto alto quanto quello del null model per il quale tutte le stime dei coefficienti sono 0 ($\lambda = \infty$). Però per un valore intermedio di λ l'MSE è considerevolmente più basso.

VANTAGGI COMPUTAZIONALI

> 💡 Ridge regression ha un significativo vantaggio computazionale rispetto a best subset selection.

Per ogni dato λ c'è bisogno di fare il fit di un solo modello e le computazioni sono molto semplici.

Risulta possibile dimostrare (non lo faremo) che le computazioni richieste per stimare i coefficienti di ridge regression simultaneamente per tutti i valori di λ, sono quasi identiche a quelle necessarie per fare il fit least squares del modello.

Least Absolute Shrinkage and Selection Operator (Lasso, operatore di minimo assoluto restringimento e selezione)

La ridge regression, differentemente da subset selection che seleziona modelli che coinvolgono solo un sottoinsieme delle variabili, ha un ovvio svantaggio, ridge regression includerà tutti i p predittori nel modello finale.

Lasso è una alternativa relativamente recente a ridge regression che supera questo svantaggio. Il coefficiente Lasso, $\hat{\beta}_\lambda^L$, minimizza la quantità

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}|\beta_j| \;= \mathrm{RSS} + \lambda\sum_{j=1}^{p}|\beta_j|.$$

In termini statistici, lasso usa una penalità l_1 (ell one o one norm) invece di una penalità l_2 (norma euclidea).

> 💡 La norma l_1 di un vettore di coefficienti β è dato da $||\beta||_1 = \sum|\beta_j|$.

Come per la ridge regression, lasso restringe le stime dei coefficienti verso 0, però la penalità l_1, usata da lasso, ha l'effetto di forzare alcune delle stime dei coefficienti ad essere esattamente pari a 0 quando il tuning parameter λ è sufficientemente grande.

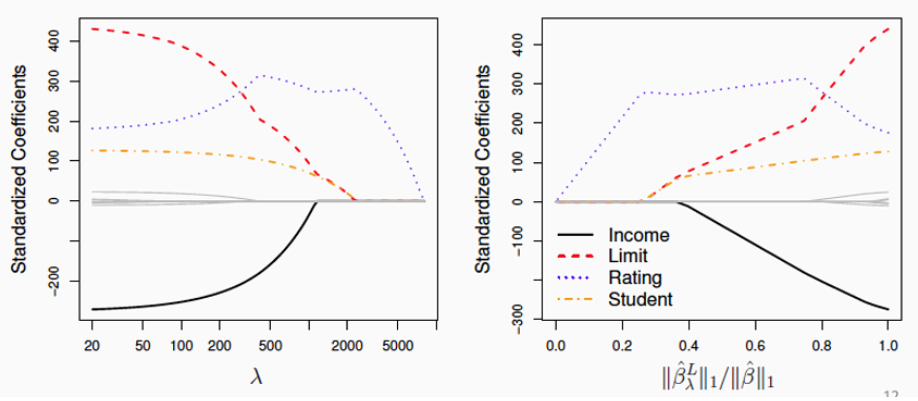
Di conseguenza, come best subset selection, lasso effettua la variable selection.

Si dice che lasso restituisce **modelli sparsi**, cioè modelli che coinvolgono solo un sottoinsieme delle variabili.

Come in ridge regression, selezionare un buon valore di λ per lasso è critico, il metodo di scelta è ancora una volta cross-validation.

Credit card data: lasso

Di seguito sono mostrati i coefficienti ridge regression standardized.



La notazione $||\beta||_1 = \sum_{j=1}^{p} |\beta_j|$ è la norma l_1 di β.

Sulla base del valore di λ, lasso può produrre un modello che coinvolge solo un sottoinsieme delle variabili.

Quando $\lambda = 0$ lasso restituisce il fit least squares, mentre per λ sufficientemente grande lasso restituisce il null model che ha tutte le stime dei coefficienti pari a 0.

Formulazioni alternative

Formulazioni alternative per Ridge regression e lasso

Un modo equivalente di scrivere il problema della ridge regression è minimizzare

$$\text{RSS} = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \text{ subject to } \sum_{j=1}^{p} \beta_j^2 \leq s.$$

Una formulazione equivalente per lasso è minimizzare

$$\text{RSS} = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \quad \text{subject to} \quad \sum_{j=1}^{p} |\beta_j| \leq s.$$

C'è una corrispondenza one-to-one tra λ nelle formulazioni precedenti ed s in quelle presentate ora.

Formulazione alternativa per best subset selection

- Consider the problem

It considers all $\binom{p}{s}$ models containing s predictors

$$\text{RSS} = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \quad \text{subject to} \quad \sum_{j=1}^{p} I(\beta_j \neq 0) \leq s.$$

Questa formulazione consiste nel trovare un set di stime di coefficienti tale che l'RSS è minimizzato, condizionatamente al fatto che non più di s coefficienti possono essere non-zero. Questo è equivalente a best subset selection.

> 💡 Visto che risolvere il problema presentato è computazionalmente infattibile quando p è grande, possiamo interpretare la ridge regression e lasso come alternative computazionalmente fattibili a best subset selection.

La proprietà di Variable Selection di Lasso

Lasso è molto più legato a best subset selection in quanto può effettuare variable selection, differentemente da ridge regression.

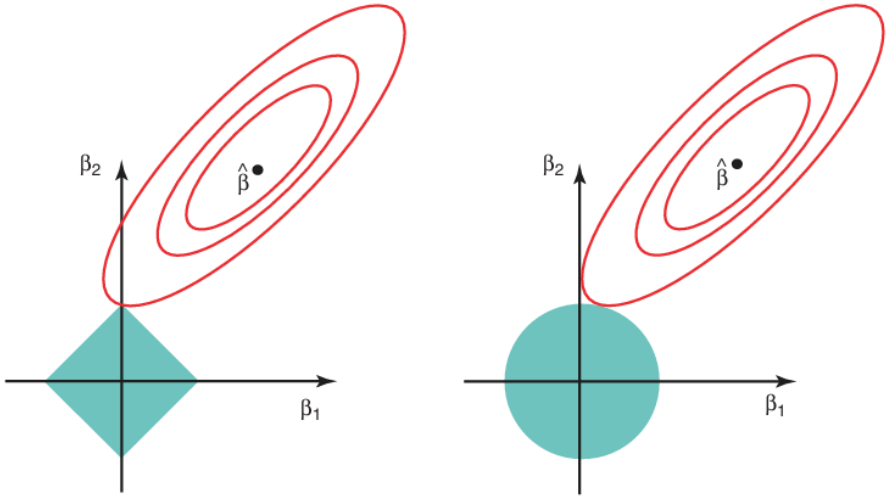
In figura sono visibili i contours dell'errore e funzioni di constraint per lasso (a sinistra) e ridge regression (a destra).

Le aree verdi sono le regioni di constraint, $|\beta_1| + |\beta_2| \leq s$ e $\beta_1^2 + \beta_2^2 \leq s$.

Le ellissi rosse sono i contours dell'RSS.

▼ Definizione di Contours

I **contorni** (o **contours**) sono linee o curve che rappresentano punti di uguale valore di una funzione in uno spazio a due dimensioni. Ad esempio, in una mappa topografica, i contorni collegano punti con la stessa altitudine, mentre in una funzione matematica bidimensionale $f(x, y)$, rappresentano i punti per cui $f(x, y) = c$, dove c è una costante. Questi sono utili per visualizzare la variazione di una funzione nello spazio.



La soluzione least squares è marcata con $\hat{\beta}$ mentre il rombo verde e il cerchio verde rappresentano i constraints di lasso e ridge regression nelle formulazioni alternative appena viste.

Se s è sufficientemente grande le regioni di constraint conterranno $\hat{\beta}$ e quindi la ridge regression e lasso saranno uguali alle stime least squares (un valore così grande di s corrisponde a $\lambda = 0$ nelle formulazioni originali).

Le formulazioni alternative indicano che le stime dei coefficienti di lasso e ridge regression sono date dal primo punto di contatto tra un ellisse e la constraint region. Quindi:

- visto che **ridge regression** ha una regione di limitazione circolare, senza punte, questa intersezione non accadrà su uno degli assi e quindi la ridge regression produrrà stime di coefficienti che saranno esclusivamente non-zero.
- visto che **lasso** ha una regione di limitazione con angoli su ognuno degli assi l'ellisse intersecherà spesso la constraint region su un asse, quando questo accade uno dei coefficienti sarà pari a 0 (in figura l'intersezione è per $\beta_1 = 0$, quindi il modello includerà solo β_2). Ad alta dimensionalità molte stime di coefficienti possono essere uguali a 0 simultaneamente.

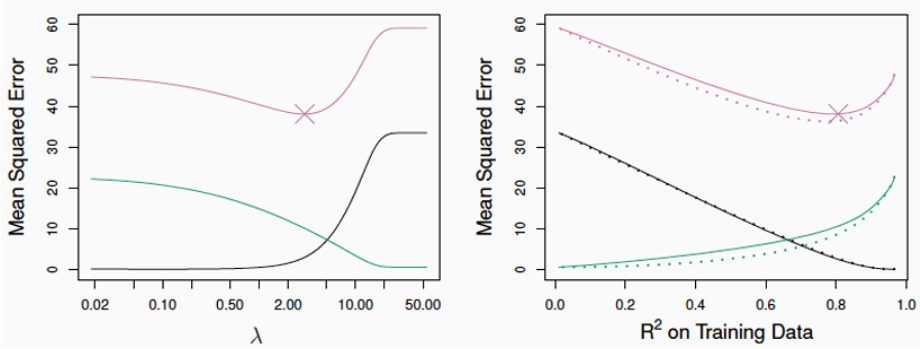
Lasso vs. ridge regression

In figura un dataset simulato con $p = 45$ ed $n = 50$.

Nero: bias. Verde: varianza. Viola: test MSE.

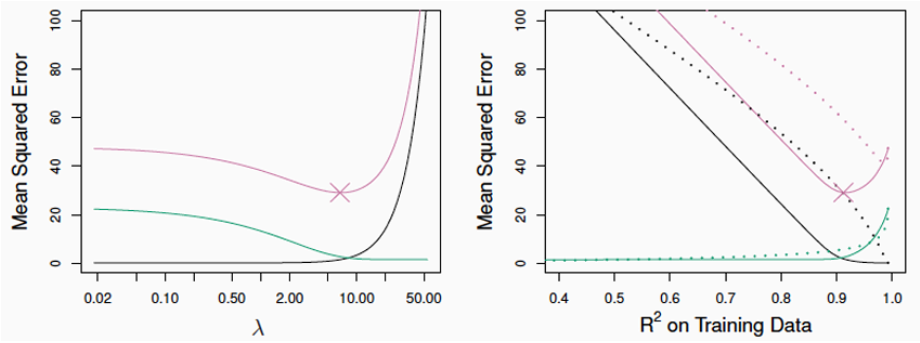
A sinistra: Lasso.

A destra: comparazione tra lasso (linee continue) e ridge regression (linee tratteggiate).



Notiamo che i bias sono quasi uguali, mentre la varianza di ridge è leggermente inferiore di quella di lasso. In questo caso ridge regression ha performato meglio di lasso. Questo è collegato al fatto che tutti i 45 predittori sono davvero legati alla response, quindi nessuno dei veri coefficienti è uguale a 0.

Nella prossima figura la situazione è differente.



In questo caso la response è funzione di solo 2 dei 45 predittori, Lasso tende a performare meglio in termini di bias, varianza ed MSE.

Nè ridge regression nè lasso supererà dominerà sull'altro universalmente.

> 💡 Lasso tende a performare meglio in un setting nel quale solo un numero relativamente piccolo di predittori ha coefficienti sostanziali.

> 💡 Ridge regression performerà meglio quando la response è funzione di molti predittori, tutti con coefficienti più o meno di pari dimensioni.

Il problema è che il numero di predittori legato alla risposta non è mai noto a priori per un vero data set. La cross-validation può essere utilizzata per determinare quale approccio è migliore su uno specifico data set.

Differentemente da ridge regression, lasso effettua variable selection, il che risulta in modelli che sono più facili da interpretare.

Discussione: best subset selection, ridge regression, lasso (ESL Ch.3)

Se X è ortonormale, le tre procedure hanno soluzioni esplicite.

Sia $\hat{\beta}_j$ la stima OLS di $\hat{\beta}$. Allora

- Best subset (dimensione s):

$$\hat{\beta}_j^B = \hat{\beta}_j \cdot I(|\hat{\beta}_j| \geq |\hat{\beta}_{(s)}|)$$

  dove $|\hat{\beta}_{(s)}|$ è l's-esimo più grande tra tutti i $|\hat{\beta}_j|_s$.

- Ridge regression:

$$\hat{\beta}^{\text{ridge}} = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y},$$

$$\hat{\beta}_j^R = \hat{\beta}_j/(1+\lambda).$$

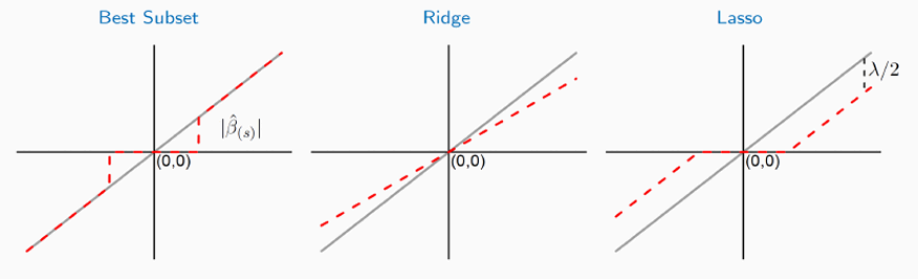
- Lasso:

$$\hat{\beta}_j^L = \text{sign}(\hat{\beta}_j)(|\hat{\beta}_j| - \lambda/2)_+,$$

  dove x_+ è uguale ad x se $x > 0$ e uguale a 0 se $x \leq 0$.

In sintesi:

- **Ridge regression** fa uno shrinkage proporzionale;
- **Lasso** traduce ogni coefficiente con un fattore costante $\lambda/2$ troncando a 0 (soft-thresholding);
- **Best subset selection** elimina tutte le variabili con coeficcienti più piccoli della s-esima più grande (questa è una forma di hard-thresholding).



La selezione del parametro di tuning λ

Come per subset selection, per ridge regression e lasso è necessario un metodo per determinare quale dei modelli in considerazione è il migliore.

Risulta quindi necessario un metodo per selezionare un valore del tuning (messa a punto, regolazione) parameter λ o, equivalentemente, un valore del constraint s.

La cross-validation fornisce un modo semplice per risolvere questo problema. Scegliamo una griglia di valori di λ e computiamo l'errore di cross-validation per ogni valore di λ.

Selezioniamo poi il valore del tuning parameter per il quale l'errore di cross-validation è il più piccolo.

> 💡 Infine, il modello è re-fitted usando tutte le osservazioni disponibili e il valore selezionato del tuning parameter.

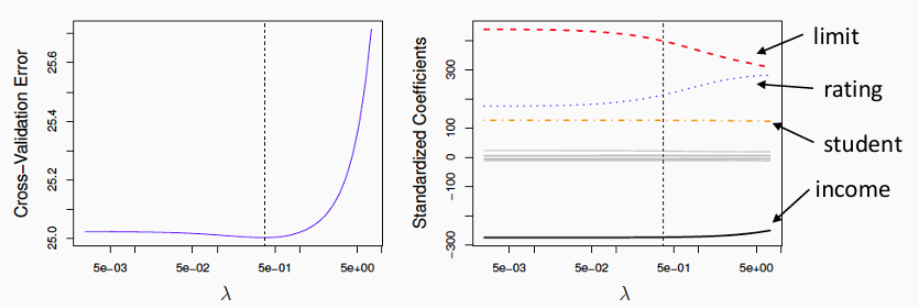
Credit data: selezione del parametro di tuning λ

Primo esempio

La figura mostra gli errori di cross-validation che vengono dall'applicazione di ridge regression a Credit data.

A sinistra: gli errori di cross-validation che risultano dall'applicazione di ridge regression al data set Credit data con vari valori di λ.

A destra: le stime dei coefficienti come funzione di λ. La linea verticale tratteggiata indica il valore di λ selezionato dalla cross-validation.



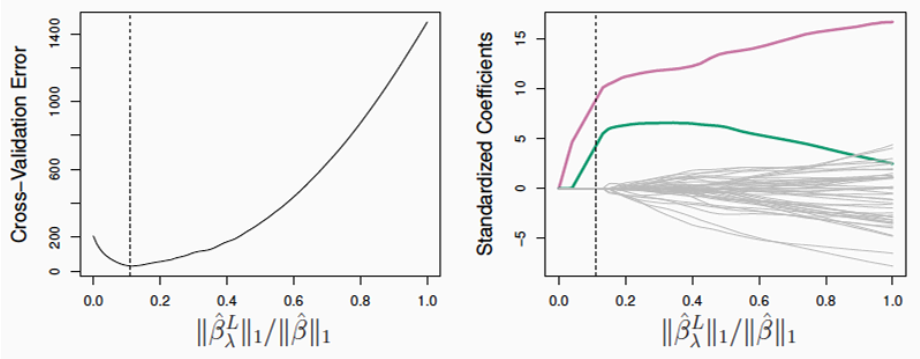
Notiamo che il λ scelto è abbastanza piccolo, il che indica una piccola quantità di shrinkage relativamente ad OLS. Inoltre la dip (il calo) dell'errore non è molto pronunciata quindi potremmo semplicemente usare OLS.

Secondo esempio

A sinistra: il 10-fold cross-validation MSE per lasso, applicato al data set simulato sparso di questa figura, dove solo 2 dei 45 predittori hanno coefficienti non-zero.

A destra: sono mostrate le corrispondenti stime lasso dei coefficienti. La linea verticale tratteggiata indica il fit lasso per il quale l'errore di cross-validation è più basso.

Le linee grigie rappresentano i predittori non relazionati alla response.



Le due linee colorate nel panel a destra rappresentano i due predittori che sono legati alla risposta, mentre le linee grigie rappresentano i predittori non legati alla risposta; ci si riferisce spesso a questi rispettivamente come signal variables e noise variables.

> 💡 Lasso da, correttamente, stime dei coefficienti più grandi ai signal predictors, ma anche il minimum cross-validation error corrisponde ad un set di stime dei coefficienti per il quale solo le signal variables sono non-zero.

Quindi, cross-validation insieme a lasso ha correttamente identificato le due variabili signal nel modello nonostante il setting complicato, con p=45 regressori e sole n=50 osservazioni.

In constrato, la soluzione least-squares (che figura nelle parte più a destra del panel di destra) assegna un coefficiente grande a solo una delle due signal variables.

Altri metodi di regolarizzazione

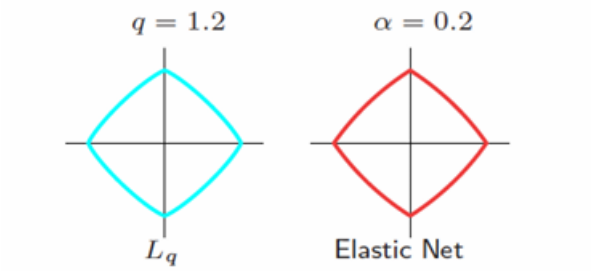
Elastic net regression

Si tratta di un metodo di regressione regolarizzato che combina linearmente le penalità l_1 ed l_2 di lasso e ridge regression.

La elastic net penalty è:

$$\lambda \sum_{j=1}^{p} \left(\alpha \beta_j^2 + (1-\alpha)|\beta_j|\right)$$

Elastic-net seleziona variabili come lasso e restringe insieme i coefficienti dei predittori correlati come ridge regression. Inoltre fornisce un vantaggio computazionale.



A sinistra i contours dei valori costanti di $\sum_j |\beta_j|^q$ per $q = 1.2$.

A destra la elastic-net penalty $\sum_j (\alpha\beta_j^2 + (1-\alpha)|\beta_j|)$ per $\alpha = 0.2$.

Nonostante siano visivamente molto simili, il plot relativo a elastic-net presenta angoli appuntiti (non-differenziabili), mentre la penalità per $q = 1.2$ non ne ha.

Documentazione `glmnet`:

> `alpha` The elasticnet mixing parameter, with $0 \leq \alpha \leq 1$. The penalty is defined as
> $$(1-\alpha)/2||\beta||_2^2 + \alpha||\beta||_1.$$
> `alpha=1` is the lasso penalty, and `alpha=0` the ridge penalty.

Least Angle Regression (LAR)

Si tratta di un metodo recente che può essere considerato una versione di forward stepwise regression.

LAR è intimamente connesso a Lasso, e infatti fornisce un algoritmo estremamente efficiente per la computazione dell'intero lasso path (percorso lasso, la sequenza di stime dei coefficienti ottenuta al variare del parametro di tuning).

And so on

Negli ultimi anni sono stati proposti molti altri metodi di regolarizzazione alternativi. Tecniche che forniscono **sparsity** (penso intenda che permettono di inquadrare correttamente situazioni nelle quali ci sono tanti predittori ma solo pochi sono veramente legati all'uscita) rappresentano ad oggi un'area di ricerca florida.