Apprendimento statistico e machine learning

Metodi di regolarizzazione: regressione Ridge e LASSO

Leonardo Egidi

Ottobre 2024

Università di Trieste

Metodi di regolarizzazione

Metodi di shrinkage

Regressione ridge e *Lasso*

- I metodi di selezione di tipo *subset* usano i minimi quadrati per stimare un modello lineare che contiene un sottoinsieme di predittori.
- Alternativamente, possiamo stimare un modello contenente tutti e p i predittori usando una tecnica che *vincola* o *regolarizza* le stime dei coefficienti, o, equivalentemente, che *restringa (shrinks)* i coefficienti verso zero.
- Potrebbe non essere subito immediato afferrare perché tale regolarizzazione dovrebbe migliorare la stima, ma risulta che *restringere* i coefficienti può significativamente ridurre la loro varianza.

Ridge regression

- Ricordiamo che la procedura dei minimi quadrati ordinari stima i coefficienti $\beta_0, \beta_1, \ldots, \beta_p$ usando i valori che minimizzano la somma dei residui al quadrato:

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

- I coefficienti delLa regressione di tipo ridge $\hat{\beta}^R$ sono i valori che invece minimizzano:

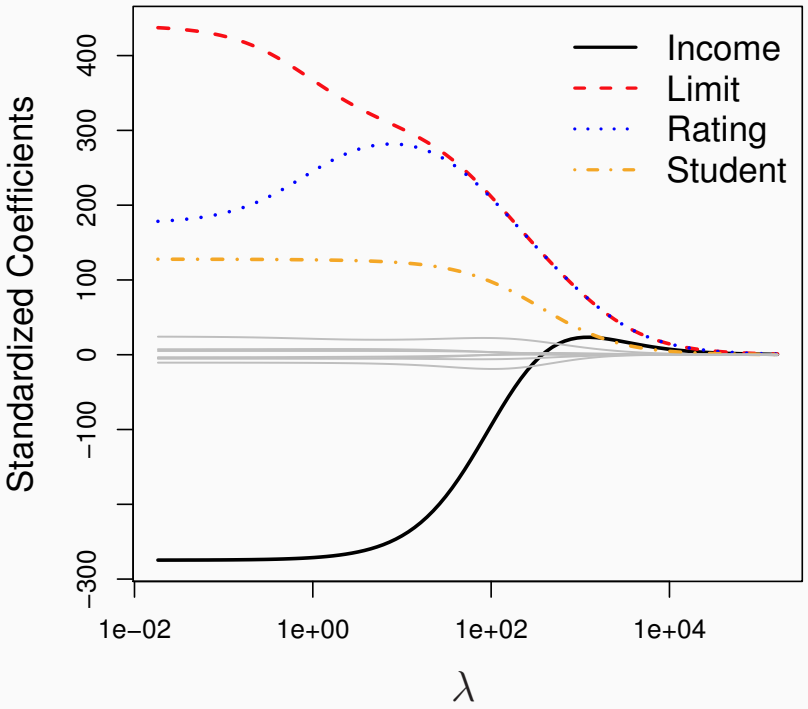
$$\sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 = \text{RSS} + \lambda \sum_{j=1}^{p} \beta_j^2,$$

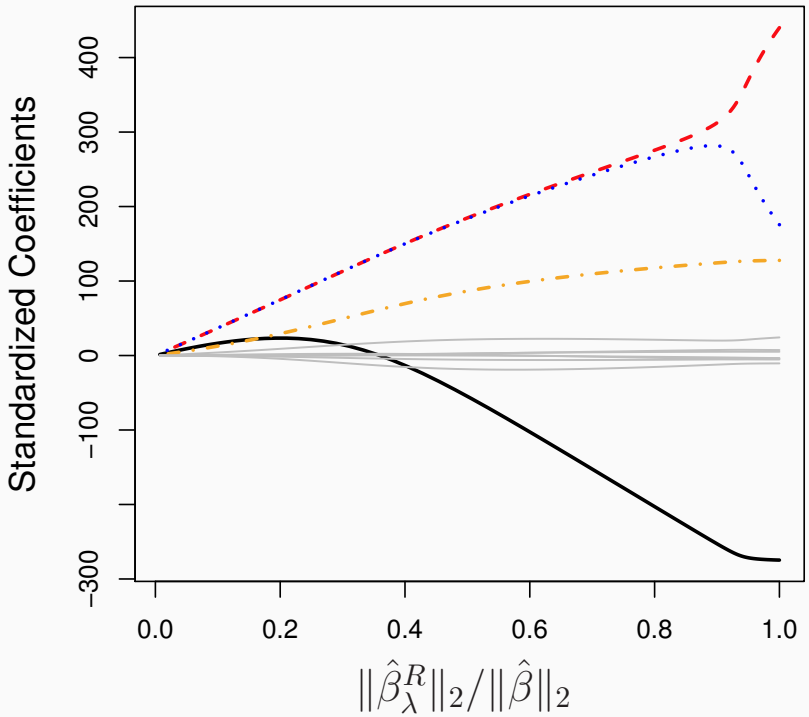
dove $\lambda \geq 0$ è un *parametro di tuning*, da stimare separatamente in un secondo momento (vedremo dopo).

Ridge regression

- Come per i metodi quadrati ordinari, la regressione ridge va alla ricerca di stime di coefficienti che si adattino bene ai dati e rendano la RSS piccola.
- Tuttavia il secondo termine $\lambda \sum_{j=1}^{p} \beta_j^2$, detto *penalità shrinkage*, è piccolo quando $\beta_1, \beta_2, \ldots, \beta_p$ sono vicino a zero, e ha quindi l'effetto di restringere le stime di β_j verso zero.
- Il parametro di tuning λ serve a controllare l'ammontare di *shrinkage (restringimento)* dei coefficienti di stima. Precisamente:
  - $\lambda = 0$ vuol dire che la penalità non ha effetto, e le stime ridge sono uguali a quelle dei minimi quadrati ordinari.
  - $\lambda \to \infty$ vuol dire che l'impatto della penalità cresce, e le stime ridge convergeranno su zero.
- Selezionare un buon valore per λ è critico: tipicamente per questo scopo viene usata la cross-validation.

Dati sui crediti





Dettagli dei grafici precedenti

- Nel grafico di sinistra, ogni curva rappresenta la stima di un coefficiente della regressione ridge per una delle dieci variabili, in funzione di λ.
- Il grafico di destra rappresenta la stessa stima dei coefficienti della ridge regression, ma invece di esserci λ, sull'asse x, c'è il rapporto $||\hat{\beta}^R_\lambda||_2/||\hat{\beta}||_2$, dove $\hat{\beta}$ indica la stima dei coefficienti secondo i minimi quadrati ordinari.
- La notazione $||\beta||_2$ indica la norma ℓ_2 di un vettore, ed è definita come $||\beta||_2 = \sqrt{\sum_{j=1}^{p} \beta_j^2}$.

Ridge regression: scalare i predittori

- Le stime dei coefficienti secondo i minimi quadrati ordinari sono *equivarianti*, il che significa che se moltiplichiamo X_j per una costante c questo semplicemente comporta una scalatura nel coefficiente β_j pari a $1/c$. In altre parole, non ha importanza come il j-esimo predittore viene scalato, $X_j\hat{\beta}_j$ rimarrà lo stesso.
- Viceversa, le stime dei coefficienti secondo la ridge regression possono cambiare *sostanzialmente* quando si moltiplica un predittore con una costante, e ciò è dovuto alla presenza di una componente di penalizzazione -la somma dei quadrati dei coefficienti- all'interno della funzione obiettivo da minimizzare.
- Quindi, è auspicabile applicare la regressione ridge *dopo aver standardizzato/scalato i predittori*, usando la formula:

$$\tilde{x}_{ij} = \frac{x_{ij}}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2}},$$

di modo tale che stiano tutti sulla stessa scala.

Perché la regressione ridge migliora i minimi quadrati ordinari?

Il tradeoff bias-varianza

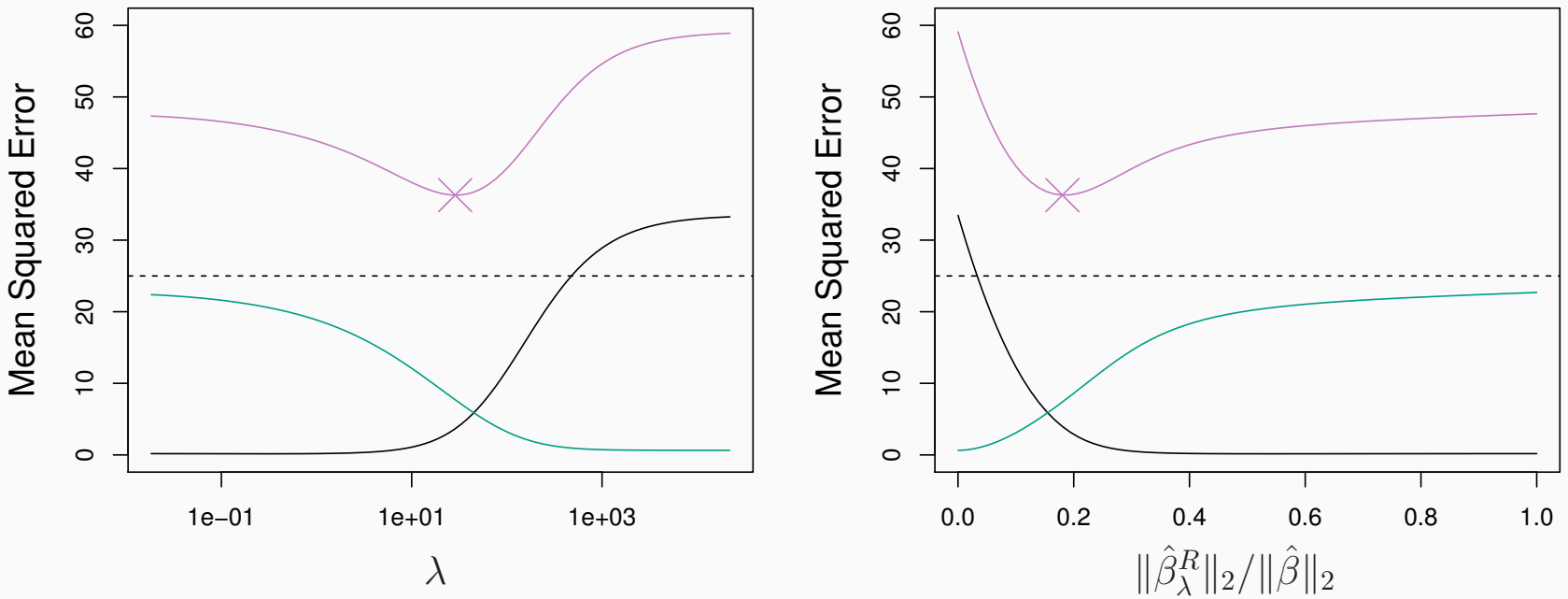


Figure 1: Dati simulati con $n = 50$ osservazioni, $p = 45$ predittori, tutti aventi coefficienti diversi da zero. Distorsione al quadrato (linea nera), varianza (linea verde) e test MSE (linea viola) per le previsioni ottenute con regresione ridge sul dataset simulato, in funzione di λ (a sinistra) e $||\hat{\beta}_\lambda^R||_2/||\hat{\beta}||_2$ (a destra). La linea orizzontale tratteggiata indica il minimo valore possibile di MSE. La croce in viola indica il modello di regressione ridge per il quale il MSE è minimo.

Lasso

- La regressione ridge ha uno svantaggio evidente: a differenza della selezione di tipo best subset, che generalmente include un sottoinsieme efficiente dei predittori, la ridge include invece tutti e p i predittori nel modello finale.
- Il *Lasso* è un'alternativa relativamente recente in grado di superare questo svantaggio. Le stime dei coefficienti secondo metodo di tipo lasso, $\hat{\beta}^L$, minimizzano la quantità:

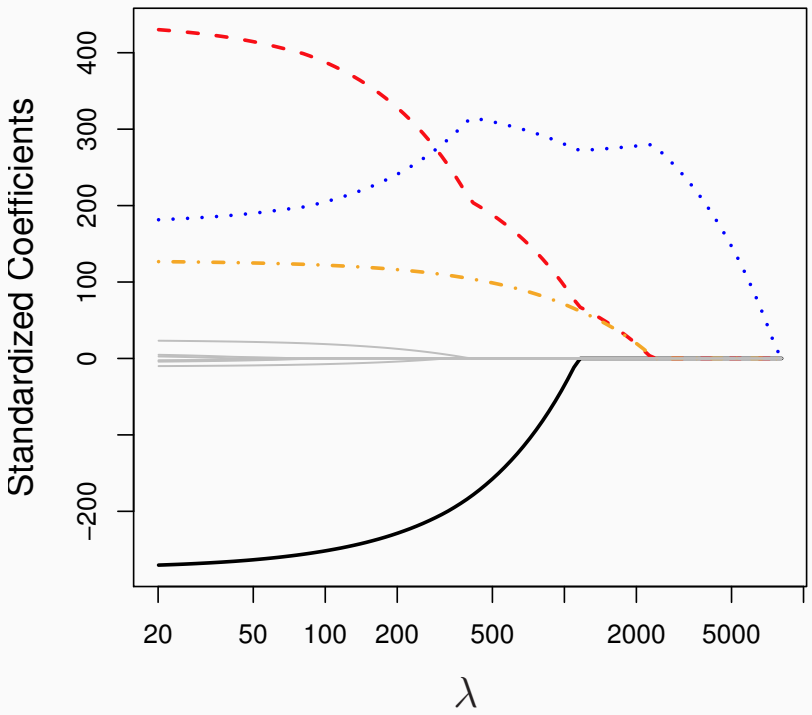
$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}|\beta_j| = \text{RSS} + \lambda\sum_{j=1}^{p}|\beta_j|.$$

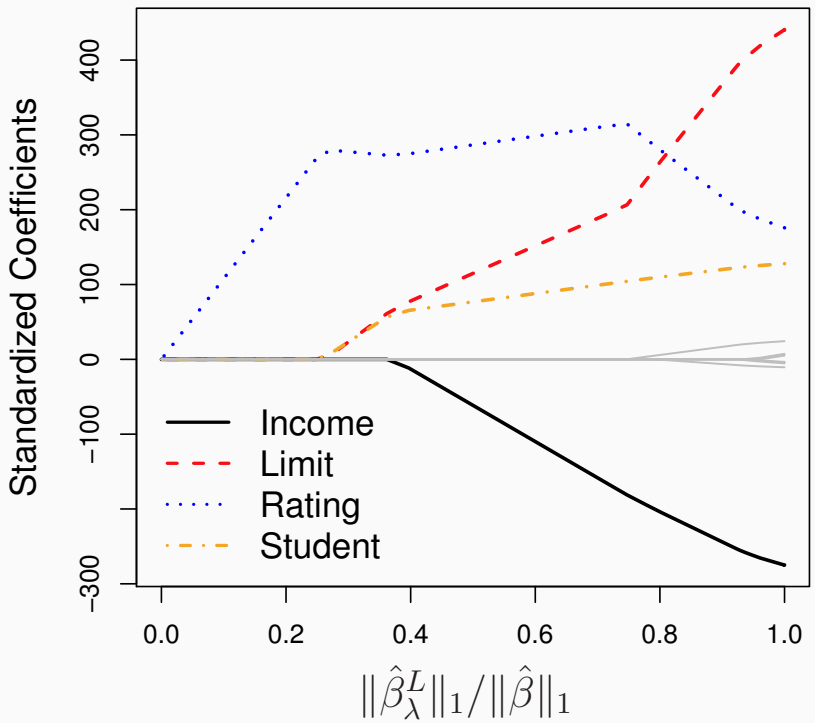
Il lasso utilizza una penalità di tipo ℓ_1 anziché di tipo ℓ_2 (usata dalla ridge). La norma ℓ_1 di un vettore di coefficienti β è definita come $||\beta||_1 = \sum_{j=1}^{p}|\beta_j|$.

Lasso

- Come nella regressione ridge, il lasso restringe le stime dei coefficienti verso zero.
- Tuttavia, nel caso del lasso, la penalità di tipo ℓ_1 forza alcune delle stime dei coefficienti ad essere esattamente uguali a zero quando il parametro di tuning λ è sufficientemente grande.
- Dunque, similarmente agli approcci best subset, il lasso fornisce una vera e propria *selezione delle variabili*.
- Diciamo che il lasso genera modelli *sparsi* - ovverossia, modelli che coinvolgono solamente un sottoinsieme delle variabili.
- Come nella regressione ridge, selezionare un buon valore per λ è uno step critico: ancora una volta ci viene in aiuto la cross-validazione.

Dati sui crediti





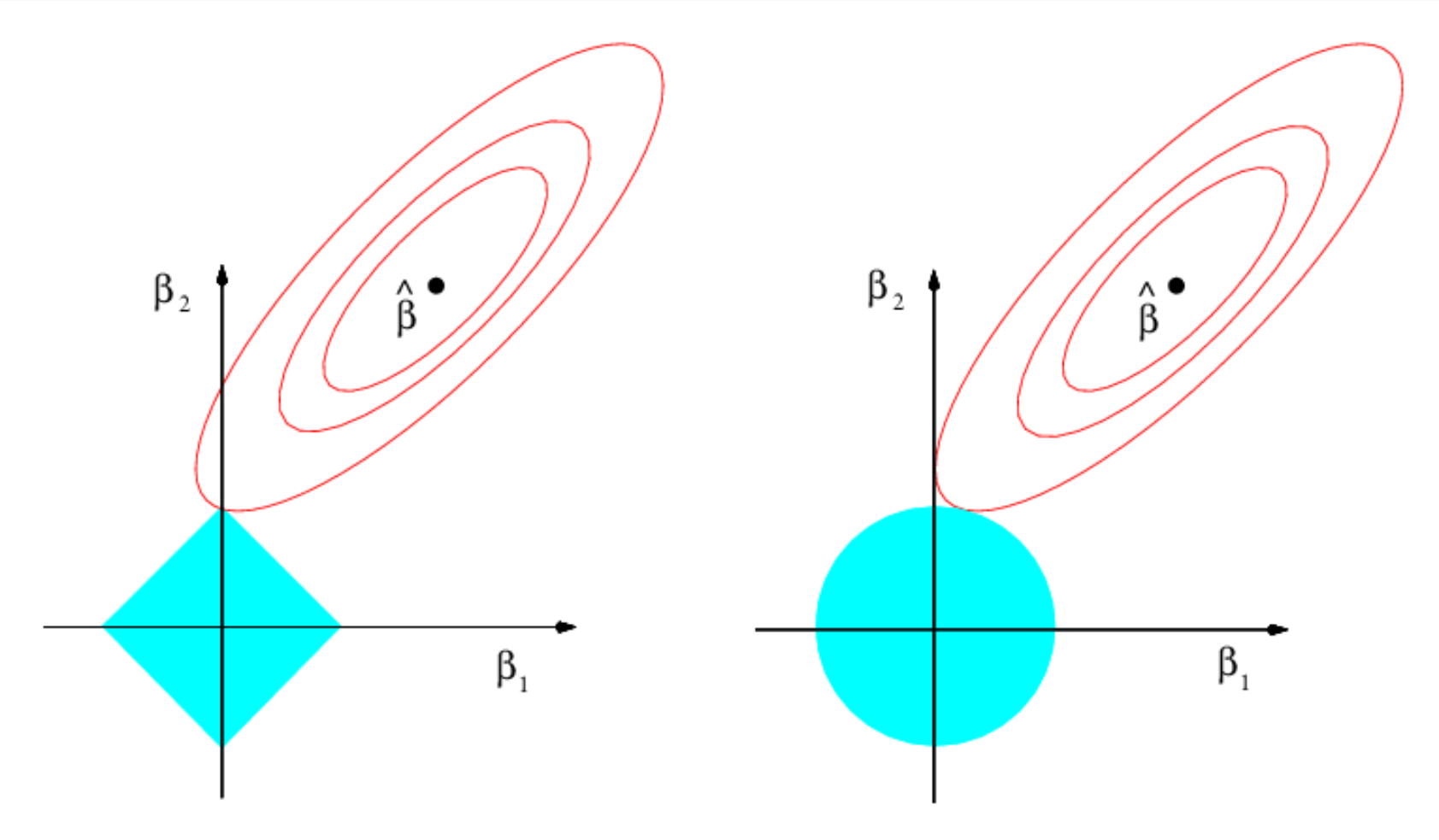
La proprietà di selezione delle variabili del lasso

- Perché il lasso conduce a stime dei coefficienti che, a differenza della regressione ridge, possono essere esattamente uguali a zero?
- Si può provare che la stima dei coefficienti secondo i metodi lasso e ridge risolvono rispettivamente i seguenti problemi di ottimizzazione vincolata:

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

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

Rappresentazione grafica di lasso e ridge



Confrontare lasso e ridge

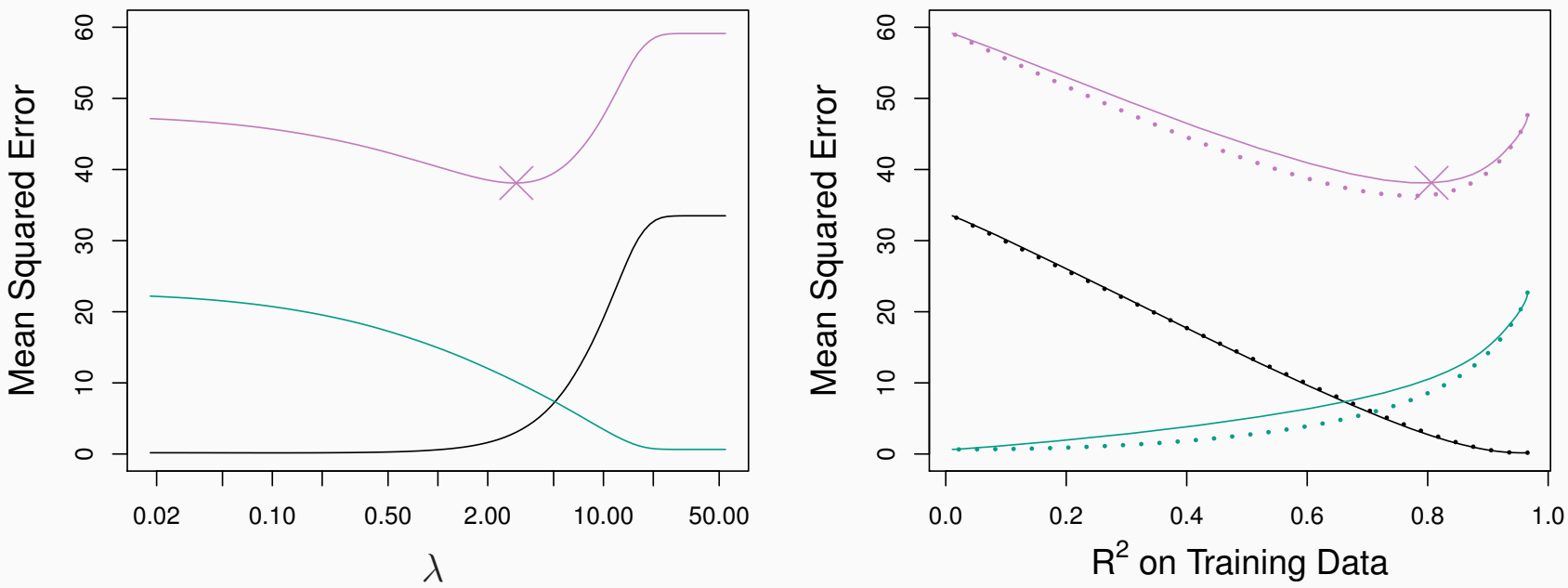


Figure 2: Sinistra: le distorsioni al quadrato (curva nera), la varianza (curva verde) e test MSE (curva viola) sui dati simulati della slide 8. Destra: confronto delle distorsioni al quadrato, varianza e test MSE tra lasso (linea solida) e ridge (linea tratteggiata). Entrambi sono plottati contro l'R^2 sui dati di training. In entrambi i grafici le croci viola indicano il modello lasso con il minore MSE.

Confrontare lasso e ridge

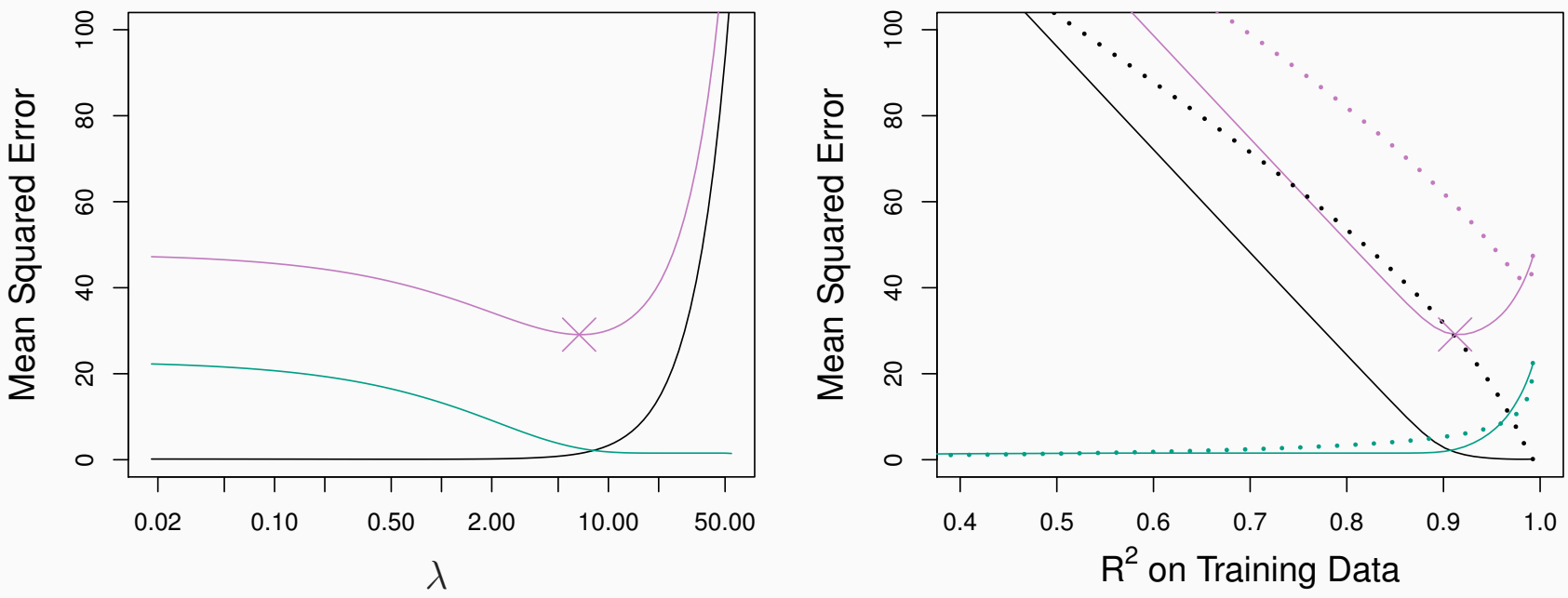


Figure 3: Sinistra: le distorsioni al quadrato (curva nera), la varianza (curva verde) e test MSE (curva viola) su dati simulati simili a quelli di slide 8, con la differenza che i predittori realmente influenti per la variabile risposta sono solamente due. Destra: confronto delle distorsioni al quadrato, varianza e test MSE tra lasso (linea solida) e ridge (linea tratteggiata). Entrambi sono plottati contro l'R^2 sui dati di training. In entrambi i grafici le croci viola indicano il modello lasso con il minore MSE.

Conclusioni

- Questi due esempi simulati sottolineano come non vi sia una chiara dominanza tra regressione di tipo ridge e lasso.
- In generale ci si può aspettare che il lasso funzioni meglio quando la variabile risposta è associata solamente a un sottoinsieme relativamente piccolo di predittori.
- Tuttavia è bene notare che il numero di predittori influenti per la variabile risposta non è mai noto a priori nei dataset reali.
- Una tecnica come la cross-validazione può essere usata per capire quale metodo vada meglio su un particolare dataset.

Selezionare il parametro di tuning λ per ridge e lasso

- Come per i metodi di tipo subset, anche per lasso e ridge ci serve un metodo per stabilire quiale dei modelli in considerazione sia il migliore.
- Ovverossia, necessitiamo di un metodo per selezionare un valore efficiente di λ o, equivalentemente, di s. (Nota bene: vi è relazione *inversa* tra λ e s. *Provatelo!*).
- La *cross-validazione* fornisce un modo semplice per raggiungere questo obiettivo. Scegliamo una griglia di valori possibili per λ e calcoliamo l'errore di cross-validazione per ognuno dei valori di λ.
- Selezioniamo poi il valore di λ in corrispondenza del quale l'errore di cross-validazione è minore.
- Alla fine il modello viene ri-stimato usando tutte le precedenti assunzioni e variabili e il nuovo valore di λ appena trovato.

Dati sui crediti

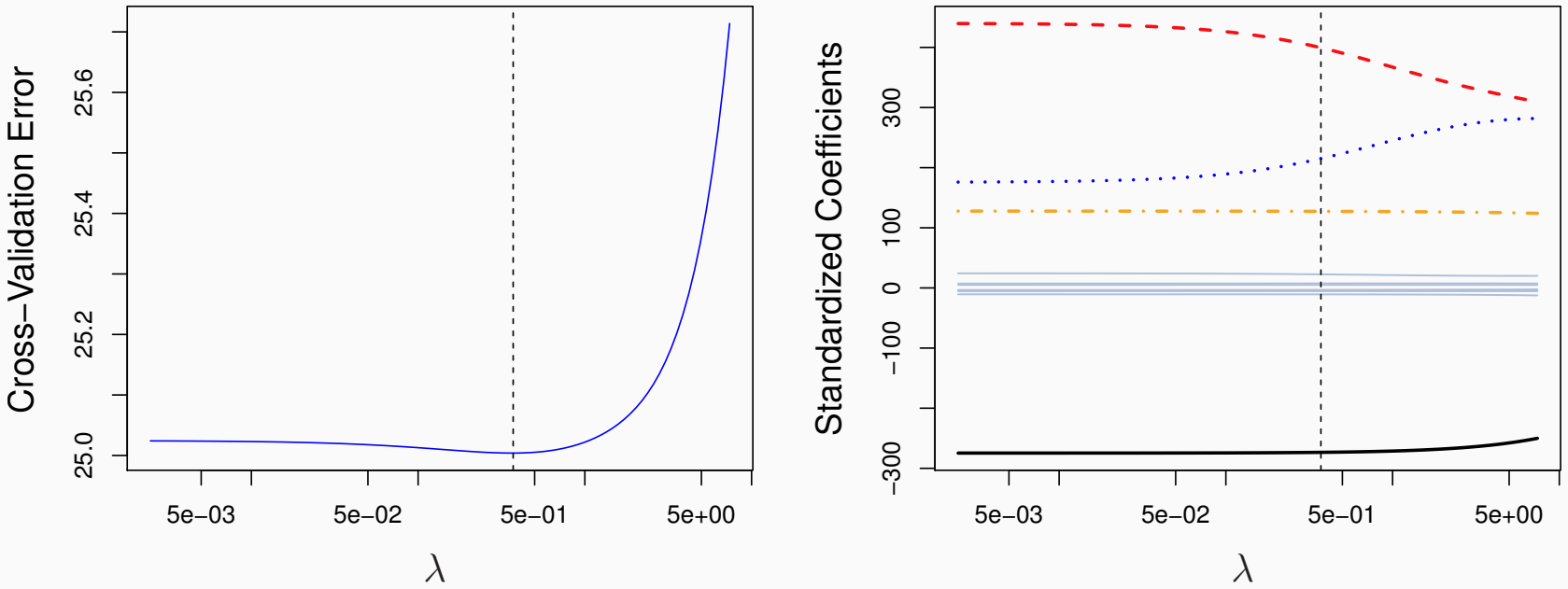


Figure 4: Sinistra: errori di cross-validazione applicando la regressione ridge ai dati `Credit` per diversi valori di λ. Destra: stime dei coefficienti in funzione di λ. La linea verticale corrisponde al valore trovato tramite cross-validazione.

Dati simulati

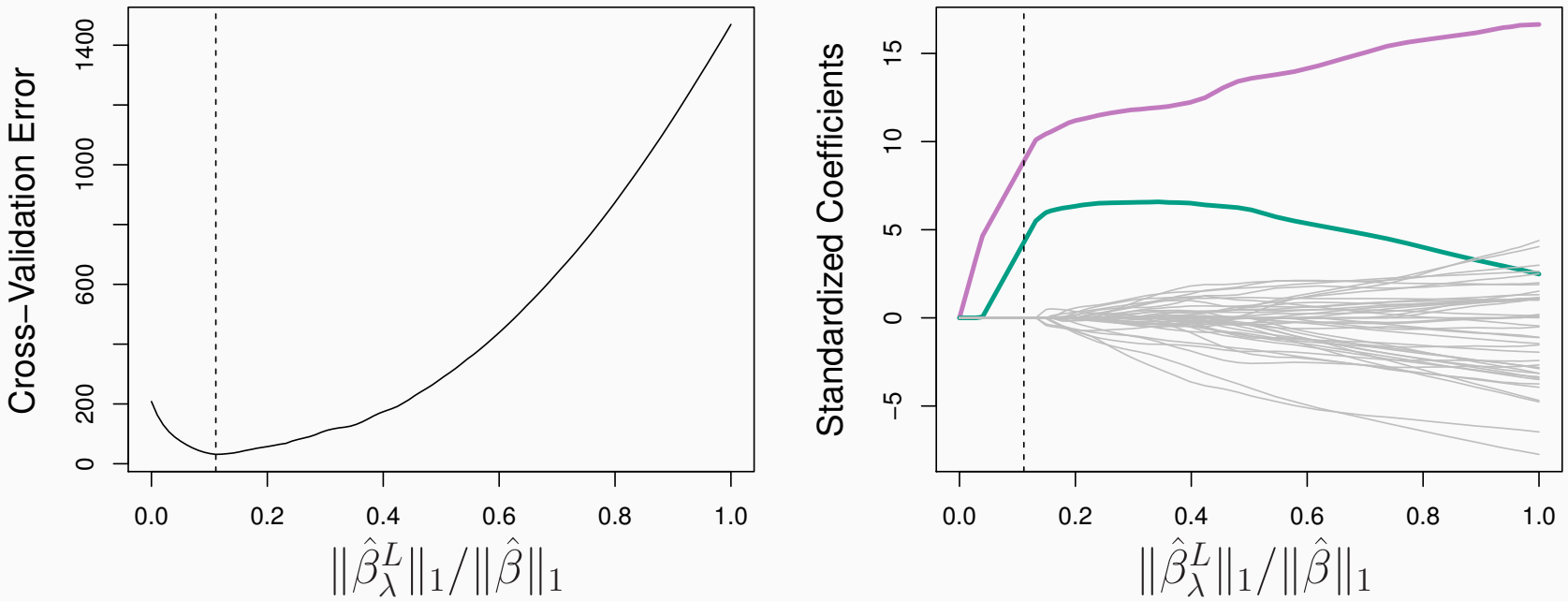


Figure 5: Sinistra: MSE con cross-validation a dieci blocchi per il lasso, applicato ai dati simulati sparsi di slide 15. Destra: stime dei coefficienti lasso corrispondenti. La linea verticale corrisponde al valore trovato tramite cross-validazione.

Commenti

- In generale, i metodi di selezione dei modelli sono uno strumento essenziale di analisi dei dati, specialmente in presenza di dataset grandi con molti predittori.
- La ricerca verso metodi che inducono *sparsità*, come il lasso, è più che mai di grande appeal.
- Ci sono approcci che tentano di mettere assieme i benefici di ridge e lasso, come ad esempio il cosiddetto *elastic net*.