11-10-2021

Seminar R / Rstudio

R Markdown → πακέτο (knitr) του R.

Τετάρτη: Εφαρμογές (και ασκήσεις)

Κάνω από την υπόθεση

$$Y|X=x = x^T b + \varepsilon, \qquad \varepsilon \sim N(0,\sigma^2)$$

$\varepsilon_1, \dots, \varepsilon_N$: ανεξάρτητα

$$\hat{b} = Gy$$

$$E\hat{b} = b \;, \quad Var(\hat{b}) = (X^T X)^{-1}\cdot\sigma^2$$

$$y \sim N(x^T b, \sigma^2 I_N)$$

$$\Rightarrow \boxed{\hat{b} \sim N\left(b, (X^T X)^{-1}\sigma^2\right)}$$

→ Διαστ. εμπιστ. b, έλεγχοι t-test

$$\boxed{(N-p-1)\hat{\sigma}^2 \sim \sigma^2 \chi^2_{N-p-1}}$$

$\hat{b}, \hat{\sigma}^2$ ανεξάρτητες

Θεώρημα Gauss-Markov: $\hat{b}$: έχουμε διαστολή ανάγεται στις γραμμικές - αμερόληπτες εκτιμήτριες του b.

Επιλογή Μεταβλητών / Μέθοδοι Συνοχής Μεταβλητών

Αν p "μεγάλο" σε σχέση με το N

$df_{er} = N - p - 1 \rightarrow$ (5p ή 10p) Εμπειρικό

$\hat{\sigma}^2 \quad \frac{\sum (y_j - \hat{\mu})^2}{N-1} \qquad Y_j \sim N(\mu, \sigma^2)$

αντί του άγνωστου $x_j^T b =$

$\hat{\sigma}^2 = \frac{\sum (y_j - x_j^T \hat{b})^2}{N-(p+1)} \qquad b \in \mathbb{R}^{p+1}$

$\rightarrow df_{error}.$

Αν έχουμε πολλές ανεξ. μεταβλητές.

Έλεγχοι υποθέσεων : $H_0: b_j = 0, \quad H_1: b_j \neq 0$

Το αποτέλεσμα του ελέγχου εξαρτάται από όλο το μοντέλο κ' όχι μόνο από τη μεταβλητή X_j

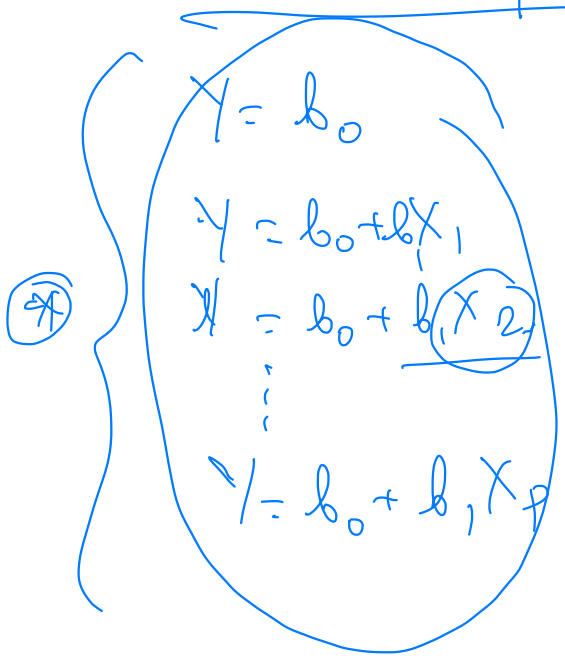
Λόγω πολυσυγγραμμικότητας

υπάρχουν αμφιβολίες ως προς τις ανεξάρτητες μεταβλητές.

Επιλογή Μεταβλητών

$X_1, \dots, X_p$ ανεξάρτητες μεταβλητές

2^p δυνατά μοντέλα



$Y = b_0 + b_1 X_1 + b_2 X_2$

$Y = b_0 + b_1 X_1 + b_2 X_3$

$Y = b_0 + b_1 X_1 + \dots + b_p X_p$

πλήρες

Best subset selection Methods

Συγκρίνουμε όλα τα δυνατά μοντέλα κι επιλέγουμε το καλύτερο με βάση κάποιο κριτήριο.

Προσεγγιστικές Μέθοδοι Επιλογής Μεταβλητών

Stepwise Regression

↙

forward

προσθέτει μια μεταβλητή σε κάθε βήμα

Άλλη Προσέγγιση

$$\begin{cases} \min RSS(b) \\ \text{s.t. αρ. μεταβ. στο μοντέλο} \le k \end{cases}$$

$$\Updownarrow$$

$$\min RSS(b) \quad \sum_{j=1}^{p} 1(b_j \neq 0) \le k.$$

$$\min RSS(b) \quad \text{s.t.} \sum_{j=1}^{p} b_j^2 \le t$$

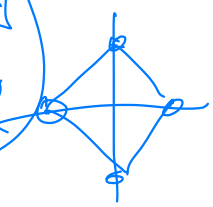
ridge regression

$$\min RSS(b) \quad \text{s.t.} \sum_{j=1}^{p} |b_j| \le t.$$

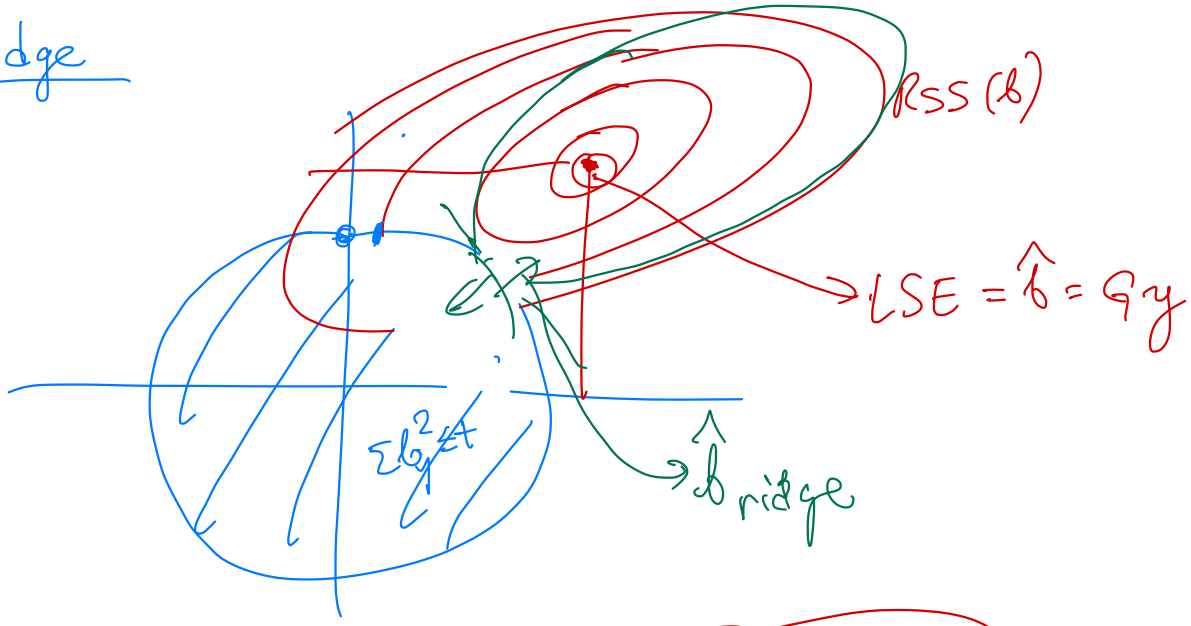
LASSO regression

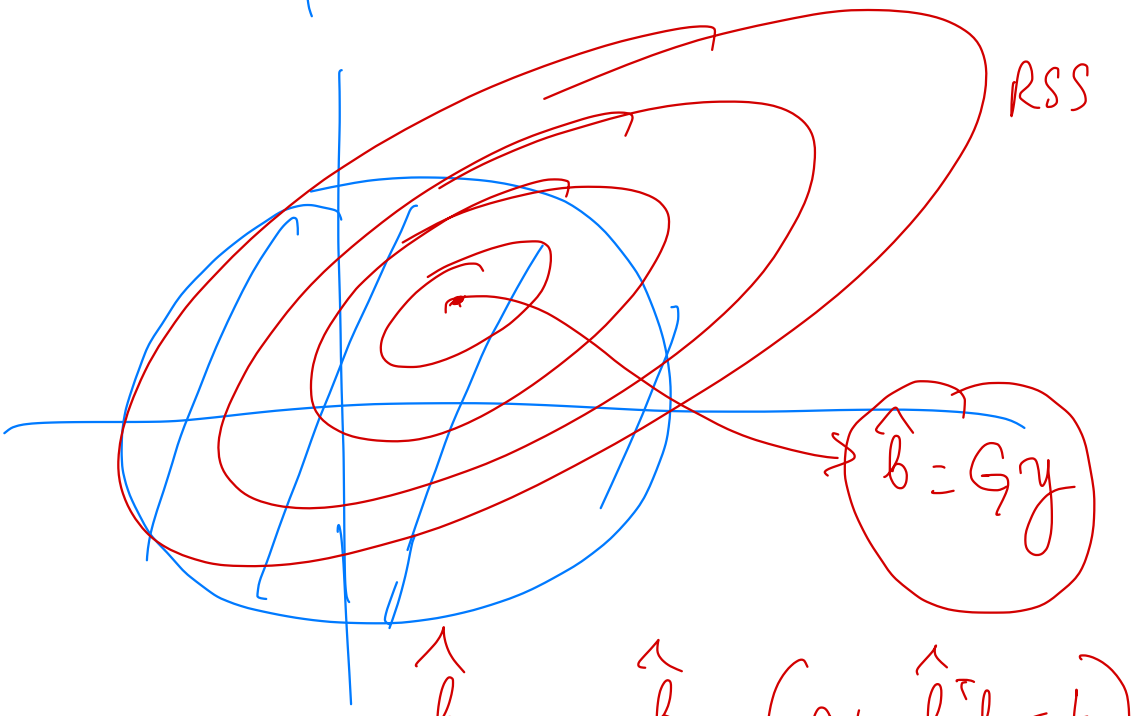
(συνορφή) Shrinkage methods

Πρόβ. κυρτού προγραμματισμού

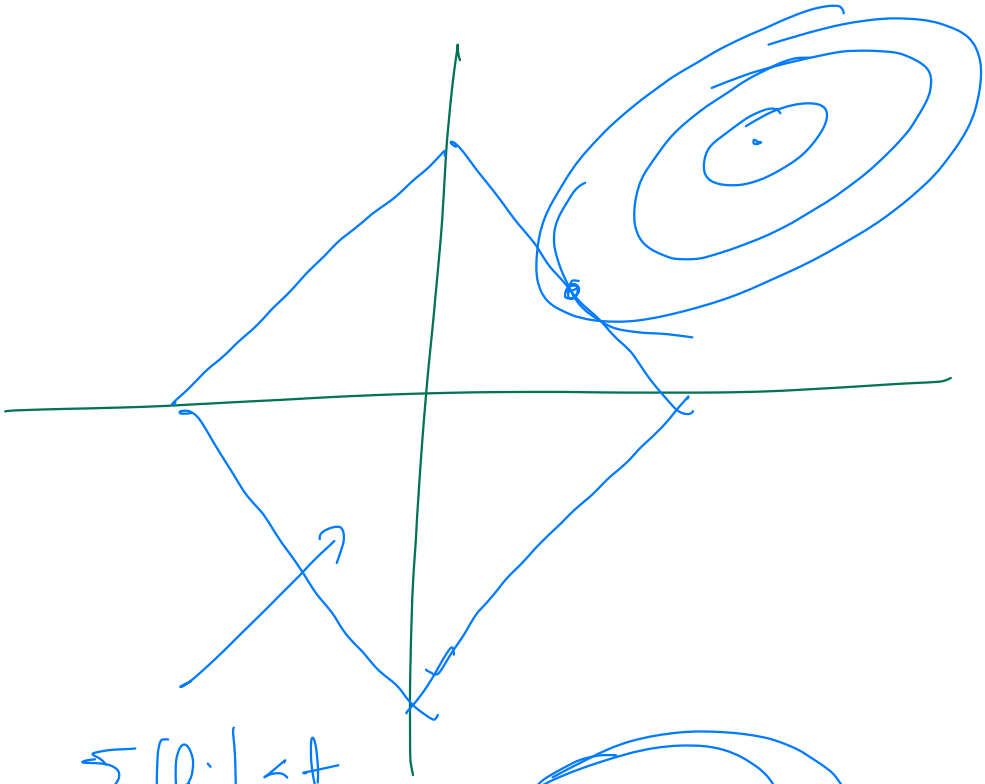


Ridge





$\hat{b}_{ridge} = \hat{b}$ (av $\hat{b}^T b \leq t$)



$\sum |b_j| = t$