Ridge & Lasso

Η Παλινδρόμηση Κορυφογραμμής (Ridge) εφαρμόζεται όταν παρουσιάζεται υψηλή συσχέτιση των επεξηγηματικών μεταβλητών (πολυσυγγραμμικότητα).

ΣΥΝΕΠΕΙΕΣ: μεγάλα τυπικά σφάλματα (standard errors) στις Ε.Ε.Τ. (LSE) και αδυναμία εντοπισμού σημαντικών και μη σημαντικών μεταβλητών.

Η Ridge (Hoerl & Kennard, 1970) συρρικνώνει τους συντελεστές παλινδρόμησης.

Οι εκτιμήτριες προκύπτουν από την ελαχιστοποίηση της

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

- $\lambda \geq 0$ = tuning parameter
- Απαιτούμε RSS=MINIMUM
- Shrinkage Penalty Term = small όταν τα $\beta \rightarrow 0$.
- $\lambda=0 \rightarrow$ E.E.T.
- $\lambda \rightarrow \infty \rightarrow$ Οι εκτιμητές $\rightarrow 0$
- Διαφορετικά λ δίνουν διαφορετικούς εκτιμητές
- Ο σταθερός όρος ΔΕΝ επηρεάζεται

- **Όταν το β μειώνεται η μείωση της διασποράς είναι προφανής. Αν πχ**

$$\beta^{*} = \frac{1}{1+\alpha}\beta$$

τότε

$$Var(\beta^{*}) = \left(\frac{1}{1+\alpha}\right)^{2} Var(\beta)$$

Πως επιδρά στην ελαχιστοποίηση του RSS, ο περιορισμός ότι το άθροισμα των τετραγώνων των β είναι μικρός $\sum \beta_i^2 \leq c$;

πώς "φαίνεται" το Ridge αποτέλεσμα;

Οι συντελεστές συρρικνώνονται προς το 0
Δεν μηδενίζονται (σε αντίθεση με Lasso)
Όσο μεγαλώνει το λ:

ο κύκλος μικραίνει
το σημείο λύσης πλησιάζει το (0,0)

ǳȘô ŰȒîŮȒǽȄòîűŭê űŭ Ůûò (RSS Ìŭê Ridge)âů řÎůűÔêÌǽ;

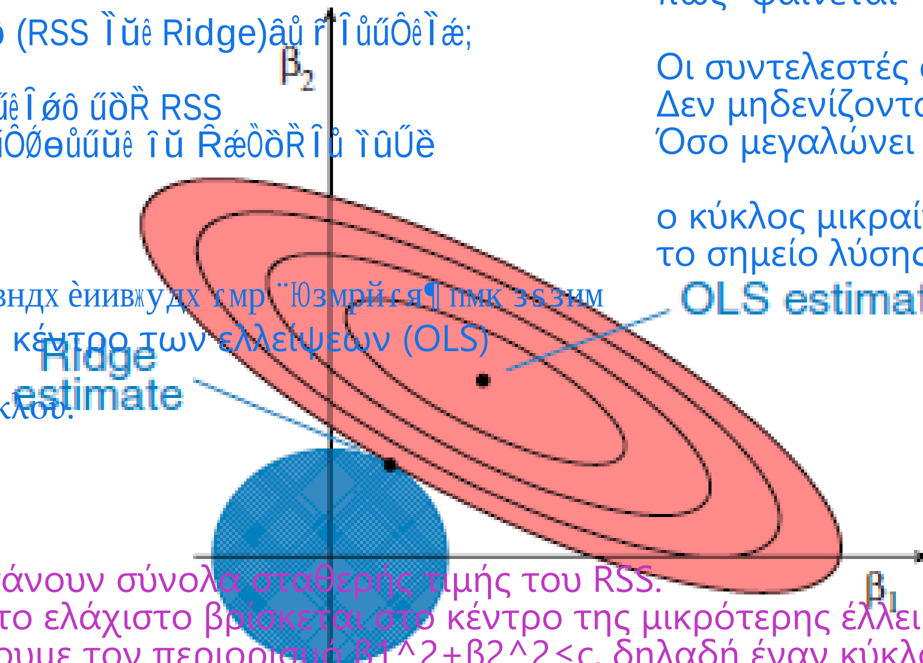
ǩìîůȀȒůêô ǚ ůɵȀɵůŮŭ ȀŰèô űêÎǿô űòȒ RSS
ǭûÎìò ǚ ɵůÔêòȉǿ ɵòȒ ůɵêűÔǿɵůűŭê îŭ ȒǽȌòȒÎů ìûŰè

λύση της Ridge είναι:

пм одйвђм вґЮсёх пдх йжзнєпвндх èиивжудх ґмр ̈Юзмрй ґя¶пмк зєзим

Δεν μπορούμε να πάμε στο κέντρο των ελλείψεων (OLS)

γιατί αυτό είναι εκτός του κύκλου.



Οι ελλείψεις παριστάνουν σύνολα σταθερής τιμής του RSS.
Χωρίς περιορισμό, το ελάχιστο βρίσκεται στο κέντρο της μικρότερης έλλειψης (OLS).
Με Ridge, όμως, έχουμε τον περιορισμό β1^2+β2^2≤c, δηλαδή έναν κύκλο.
Άρα η λύση είναι το σημείο επαφής της μικρότερης έλλειψης με αυτόν τον κύκλο

Ελλείψεις: Σε κάθε έλλειψη το RSS έχει ίδια τιμή. Οσο μικρότερη η έλλειψη τόσο μικρότερο το RSS. Το ελάχιστο στο κέντρο της μικρότερης έλλειψης.

Τώρα: Η ελαχιστοποίηση γίνεται εντός του χωρίου που ορίζει ο περιορισμός. Για p=2 ο περιορισμός είναι ο κύκλος

$$\beta_1^2 + \beta_2^2 \leq c$$

$\lambda \uparrow \rightarrow$ εκτιμητές $\downarrow$

Οπότε συρρικνώνεται η διασπορά (standard error)

Εναλλακτικά: L2 Norm $\sqrt{\sum \beta_j^2} = \| \beta \|_2$

$\lambda \uparrow \rightarrow$ Νόρμα $\downarrow$

Εναλλακτικά: Standardized Coefficients $\| \beta^{ridge} \|_2 / \| \beta \|_2$

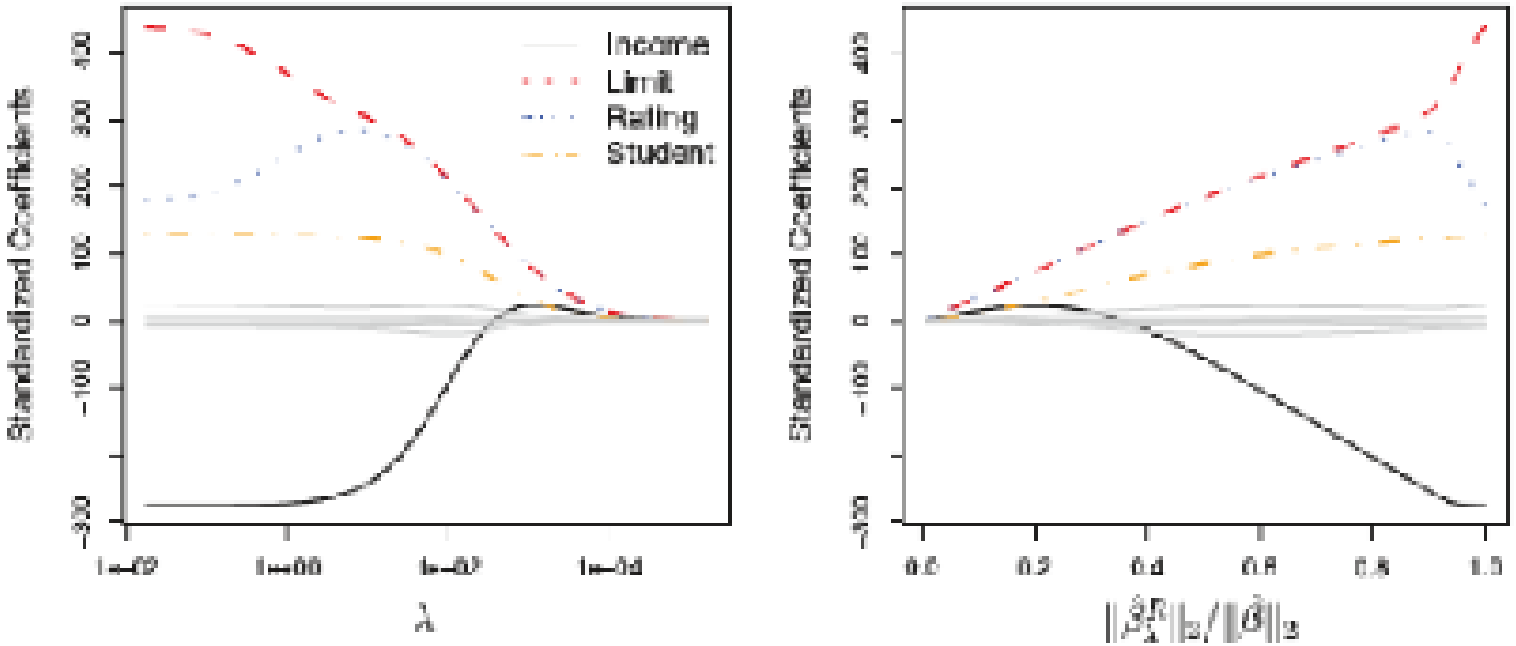
$\lambda = 0 \rightarrow$ Stand. Coeff. = 1

$\lambda \rightarrow \infty \rightarrow$ Stand. Coeff. = 0

Για λ = 0, η Ridge ταυτίζεται με τις ΕΕΤ
Οι συντελεστές ξεκινούν από τις OLS (ΕΕΤ) τιμές τους

Όσο το λ αυξάνεται:
Όλοι οι συντελεστές συρρικνώνονται
Οι "μεγάλοι" συντελεστές ↓ πιο αργά
Οι "μικροί" ↓ πιο γρήγορα
Όλοι όμως τείνουν στο 0 όταν λ→∞

Σχήμα 2.5: Γραφική παράσταση των ορθοκανονικοποιημένων μεταβλητών συναρτήσει του $\hat{\beta}$ και της ποσότητας $\|\hat{\beta}_\lambda^R\|_2 / \|\hat{\beta}\|_2$



Η Ridge μειώνει το μέγεθος όλων των συντελεστών, αλλά δεν τους μηδενίζει (γιατί χρησιμοποιεί L2 penalty)

ΥΠΟΛΟΓΙΣΜΟΣ

$$\beta^{ridge} = \min_{\beta}\{\| Y'Y - 2\beta' X'Y + \beta' X'X \beta + \lambda \| \beta \|_2\}$$

Παραγωγίζοντας έχουμε

$$\partial .../\partial\beta = -2X'Y + 2X'X\beta + 2\lambda\beta = 0$$

δηλ. X'X=I άρα-2X'Y+2β+2λβ=0

=X'XX'Y = X'Y

Αν X=ορθογώνιες τότε οι Ε.Ε.Τ. είναι X'Y και αρα $X'X\beta_{LSE} = X'Y$

$$\partial .../\partial\beta = -2\beta_{LSE} + 2\beta + 2\lambda\beta = 0$$

$$\beta_k^{ridge} = \frac{1}{1+\lambda}\beta_{LSE}$$

λ = 0 → Ridge = OLS
λ ↑ → βridge↓
λ → ∞ → βridge→0

ΥΠΟΛΟΓΙΣΜΟΣ ΔΙΑΣΠΟΡΑΣ

$$-2\mathrm{X}'\mathrm{Y}+2X'X\beta+2\lambda\beta=0\rightarrow \mathrm{X}'\mathrm{Y}-[X'X+\lambda\cdot\mathrm{I}]\beta=0$$

$$Var(\beta_k^{ridge})=Var\{[X'X+\lambda\cdot\mathrm{I}]^{-1}\mathrm{X}'\mathrm{Y}\}=$$

$$=[XX'+\lambda\cdot\mathrm{I}]^{-1}\mathrm{X}'Var(Y)X[X'X+\lambda\cdot\mathrm{I}]^{-1}$$

ΠΛΕΟΝΕΚΤΗΜΑ/ΜΕΙΟΝΕΚΤΗΜΑ:

- Όσο αυξάνεται το λ μειώνεται μεν η διασπορά ΑΛΛΑ αυξάνεται η μεροληψία (μεροληπτικές εκτιμήτριες). Αποτυπώνεται στην αύξηση της εκτίμησης του σταθερου όρου β_0.

- Συστήνεται όταν έχουμε μεγάλα τυπικά σφάλματα. Επιδιώκεται η επιλογή κατάλληλου λ ώστε να επιτευχθεί σημαντική μείωση της διασποράς αλλά μικρή αύξηση της μεροληψίας.

- Οι τιμές των εκτιμητών μειώνονται (και άρα και το τυπικό σφάλμα) ΑΛΛΑ δεν μηδενίζονται.

library(glmnet) & library(caret).

LASSO least absolute shrinkage and selection operator

Αλλαγή νόρμας → L_1

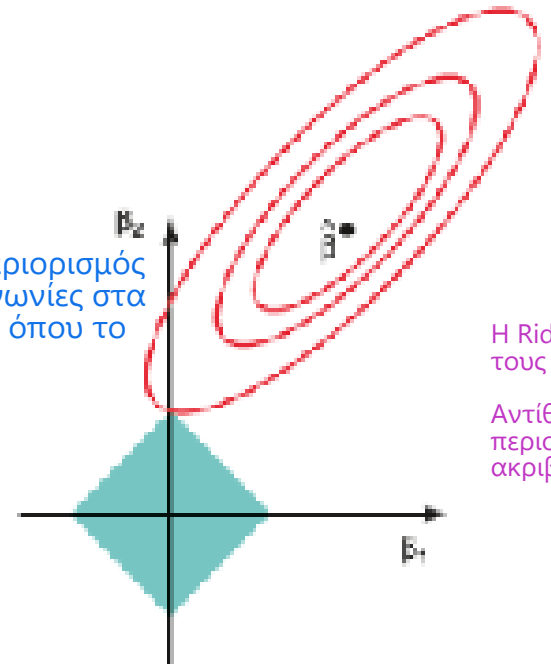
$$\sqrt{\sum |\beta|} = \|\beta\|_1$$

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

Η ελαχιστοποίηση οδηγεί κάποιους εκτιμητές συντελεστών αν γίνουν ακριβώς ίσοι με το 0.

Αρα πρόκειται για ΕΠΙΛΟΓΗ ΜΕΤΑΒΛΗΤΩΝ

Σχήμα 2.6: *Περιορισμός Lasso παλινδρόμησης σε 2 διαστάσεις*



Η Lasso μηδενίζει συντελεστές γιατί ο L1 περιορισμός (ρόμβος και όχι κύκλος τώρα) δημιουργεί γωνίες στα σημεία (±c,0), (0,±c) στο επιτρεπτό σύνολο, όπου το ελάχιστο του RSS επιτυγχάνεται συχνά.

Η Ridge χρησιμοποιεί L2 penalty, συρρικνώνει ομαλά όλους τους συντελεστές προς το μηδέν χωρίς να τους μηδενίζει.

Αντίθετα, η Lasso με L1 penalty δημιουργεί γωνίες στο χωρίο περιορισμού, με αποτέλεσμα κάποιοι συντελεστές να γίνονται ακριβώς μηδέν (επιλογή μεταβλητών)

$$|\beta_1| + |\beta_2| \leq c$$

Το σημείο καμπής είναι εκεί όπου μια από τις συντεταγμένες είναι 0.

Αλλα penalties:

$$\lambda \sum_{j=1}^{p} |\beta_j|^q$$

Elastic Net

Συνδυαστική Τεχνική Ποινικοποίησης (L1 & L2):

$\alpha = 0 \rightarrow$ Ridge
$\alpha = 1 \rightarrow$ Lasso
$0 < \alpha < 1 \rightarrow$ Elastic Net

Η Elastic Net συνδυάζει Ridge και Lasso,
προσφέροντας σταθερό shrinkage και ταυτόχρονα δυνατότητα επιλογής μεταβλητών,
ιδίως όταν υπάρχουν ισχυρά συσχετισμένοι ερμηνευτικοί παράγοντες.

$$\sum\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p} a\,|\beta_j| + \lambda\sum_{j=1}^{p}\frac{1-\alpha}{2}\beta_j^2$$

Ποιο να διαλέξω και πότε;

Ridge

Όλες οι μεταβλητές έχουν σημασία
Υψηλή πολυσυγγραμμικότητα
Θες σταθερότητα, όχι επιλογή

Lasso

Πολλές μεταβλητές, λίγες πραγματικά σημαντικές
Θες απλό, ερμηνεύσιμο μοντέλο

Elastic Net

Πολλές και συσχετισμένες μεταβλητές
Η Lasso πετάει «τυχαία» μία από ομάδα συσχετισμένων
Θες ισορροπία

Η Ridge μειώνει τη διασπορά συρρικνώνοντας όλους τους συντελεστές χωρίς να τους μηδενίζει.
Η Lasso και η Elastic Net μπορούν να μηδενίσουν συντελεστές· η Elastic Net υπερτερεί όταν υπάρχουν ομάδες συσχετισμένων μεταβλητών.

Οπτική Συγκριση

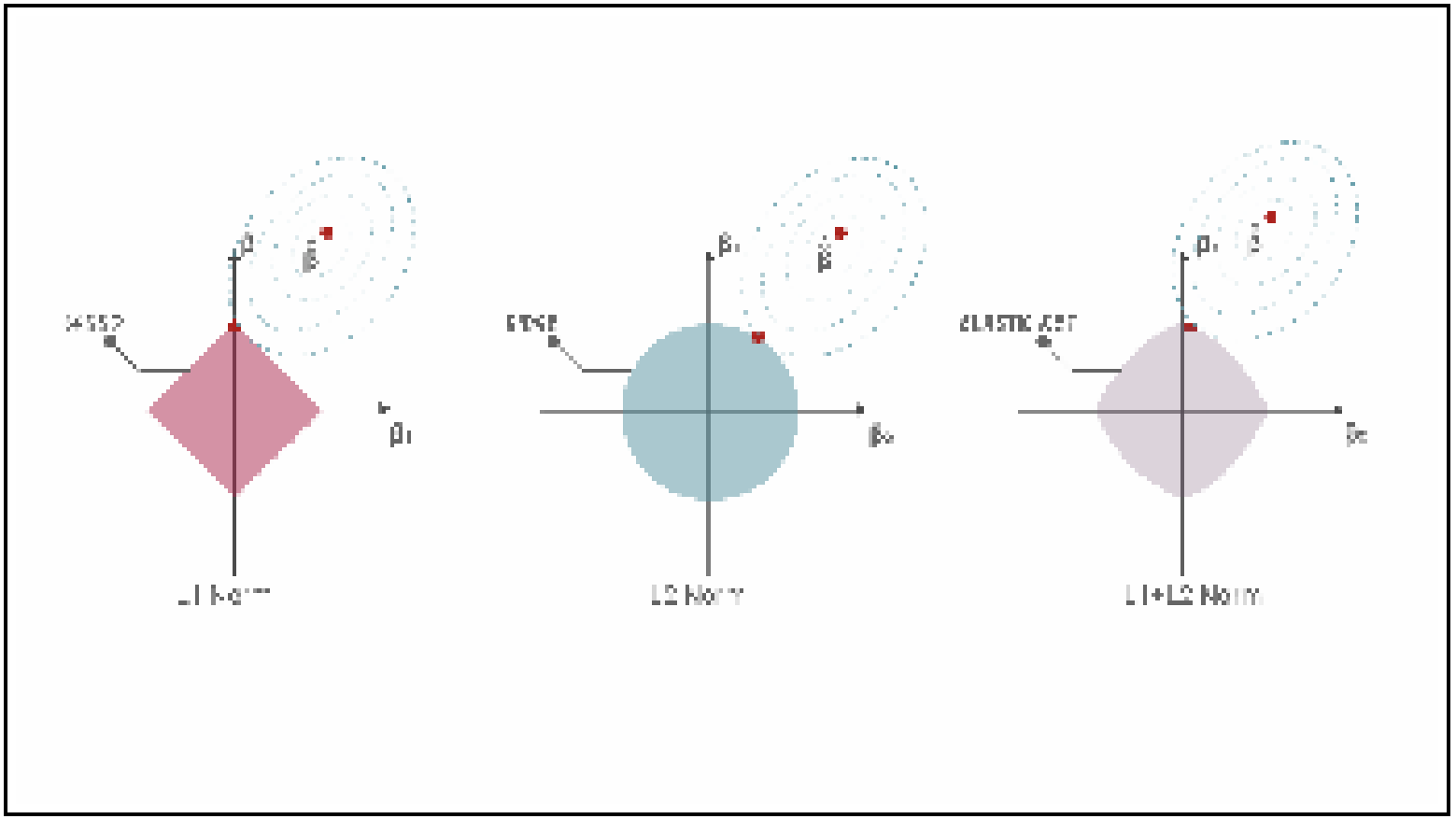


FIGURE 2.3: Geometric projection of L1 and L2 Norms' penalty terms in the space of the model parameters

Εφαρμογές

Πρόβλεψη οικονομικών δεικτών: Βελτιώνεται η πρόβλεψη οικονομικών παραγόντων/δεικτών όπως το ΑΕΠ, ο πληθωρισμός και η ανεργία, διορθώνοντας την πολυσυγγραμμικότητα μεταξύ προγνωστικών παραγόντων όπως τα επιτόκια και οι καταναλωτικές δαπάνες και έτσι οδηγώντας σε πιο ακριβείς προβλέψεις.

Ιατρική διάγνωση: Στην υγειονομική περίθαλψη, βοηθά στη δημιουργία διαγνωστικών μοντέλων ελέγχοντας την πολυσυγγραμμικότητα μεταξύ των βιοδεικτών, βελτιώνοντας τη διάγνωση και την πρόγνωση της νόσου.

Πρόβλεψη πωλήσεων: Στο μάρκετινγκ, γίνεται πρόβλεψη πωλήσεων βάσει παραγόντων όπως πχ το κόστος διαφήμισης και οι προσφορές, αξιοποιώντας συσχετίσεις μεταξύ μεταβλητών για καλύτερο σχεδιασμό πωλήσεων.

Μοντελοποίηση κλίματος: Για βελτίωση κλιματικών μοντέλων εξαλείφοντας συσχετίσεις μεταξύ μεταβλητών όπως η θερμοκρασία και η βροχόπτωση και έτσι διασφαλίζοντας πιο ακριβείς προβλέψεις.

Διαχείριση Κινδύνων: Στην αξιολόγηση πιστοληπτικής ικανότητας και στην ανάλυση χρηματοοικονομικού κινδύνου, αξιολογείται η πιστοληπτική ικανότητα αντιμετωπίζοντας την πολυσυγγραμμικότητα μεταξύ των χρηματοοικονομικών δεικτών, ενισχύοντας την ακρίβεια στη διαχείριση κινδύνου.

ΕΝΤΟΛΕΣ R

- ridge + lasso: **glmnet** + **cv.glmnet** (glmnet).

Επίσης

- linear model με ridge regression: **lm.ridge** (MASS).
- linear model με lasso: **lars** + **cv.lars** (lars).
- Penalized regression (lasso and ridge) με cross-validation routines: (**penalized**).

Penalized regression είναι η παλινδρόμηση όπου η ελαχιστοποίηση του RSS συνοδεύεται από έναν όρο ποινής στους συντελεστές, με στόχο τη μείωση της διασποράς, την αποφυγή υπερπροσαρμογής και, σε κάποιες περιπτώσεις, την επιλογή μεταβλητών.

Penalized regression = OLS + τιμωρία στο μέγεθος των συντελεστών.

Non-negative garrote

Πρόσθετο πρόβλημα: Όταν το **πλήθος των μεταβλητών είναι πολύ μεγάλο** παρουσιάζεται το πρόβλημα της επιλογής εκείνων των μεταβλητών που θα συμμετέχουν στο μοντέλο, μπορεί να υπάρχουν **πολλοί συνδυασμοί «κατάλληλων» μεταβλητών.**

Οι **Ε.Ε.Τ**. είναι αμερόληπτες που μπορεί να **υστερούν σε ακρίβεια** και έτσι η πρόβλεψη μπορεί να έχει μεγάλη απόκλιση από την πραγματική τιμή.

Η **ακρίβεια πρόβλεψης μπορεί να βελτιωθεί συρρικνώνοντας τους συντελεστές** της παλινδρόμησης ή θέτοντάς τους ίσους με το μηδέν.

Type text here

Non-negative garrote

Η τεχνική (**Breiman, 1995, Technometrics, vol. 37(4), 373-384**) επιβάλει ένα penalty για να αυξήσει την ακρίβεια. Οι νέοι εκτιμητές non-negative garrote είναι

$$\beta_{j,G} = c_j \beta_j, \quad j = 0,1,\ldots, p$$

Type text here

$c_j = παραμετρος\ συρρικνωσης$

η οποία επιλέγεται ελαχιστοποιώντας την

$$\sum \left(y_i - \sum_{j=0}^{p} c_j \beta_j x_{ij} \right)^2$$

Υπό τους περιορισμούς $c_j \geq 0 \quad \& \quad \sum_{j=0}^{p} c_j \leq s$

Non-negative garrote

Αν τα διανύσματα των επεξηγηματικών είναι ορθογώνια οπότε ο πίνακας $X'X=I$ τότε η ελαχιστοποίηση γίνεται με τη συνθήκη

$$\sum_{j=0}^{p} c_j = s \qquad (*)$$

και

$$c_j = \left(1 - \frac{\lambda^2}{\beta_j^2}\right)^+$$

με το λ να καθορίζεται από τη συνθήκη (*).

Non-negative garrote

Όσο μικρότερο το s τόσο μικρότερες τιμές λαμβάνουν οι εκτιμητές.

ΔΕΝ αφαιρείται κάποια επεξηγηματική μεταβλητή και άρα **δεν υπάρχει απώλεια πληροφορίας**.

Ο Breiman μελέτησε πραγματικά και προσομοιωμένα δεδομένα και διαπίστωσε ότι **τα σφάλματα πρόβλεψης ήταν μικρότερα.**

Non-negative garrote

- **Η τεχνική έχει μικρότερο σφάλμα πρόβλεψης**
- **Ανταγωνίζεται την τεχνική ridge**
- **Συνήθως επηρεάζεται από το πρόσημο των Ε.Ε.Τ. & από το αν υπάρχουν υψηλές συσχετίσεις στις επεξηγηματικές μεταβλητές.**

```
library(lqa)
cv.nng
lambda.nng = list(0.1, 0.2, 0.5, 1, 1.5, 2, 3,4,5)
```

ΣΥΓΚΡΙΣΗ

Η Lasso έχει ένα πλεονέκτημα σε σχέση με την Ridge καθώς παράγει μοντέλα με μεταβλητές που είναι υποσύνολο των αρχικών μεταβλητών (variable selection) και προσφέρει εύκολη ερμηνεία των μοντέλων.

Ποια τεχνική έχει καλύτερη προβλεψιμότητα;

Αν όλες οι μεταβλητές είναι σημαντικές (όλα τα β διάφορα του 0) η Lasso μηδενίζοντας κάποιες καθίσταται υποδεέστερη της Ridge.

Αν δεν είναι όλες οι μεταβλητές σημαντικές και η Lasso τις εντοπίζει τότε υπερτερεί της Ridge.

Η ElasticNet είναι χρονοβόρα διότι τρέχει (για κάθε λ) σε όλα τα πιθανά α, ώστε αν καταλήξει.

Σε πραγματικά δεδομένα δεν γνωρίζουμε εκ των προτέρων (και ούτε εκ των υστέρων) ποιες μεταβλητές είναι σημαντικές άρα θα πρέπει να εξεταστούν και οι δύο τεχνικές.

Η Ridge συρρικνώνει κάθε συντελεστή κατά ίδια αναλογία και οδηγεί σε καλύτερη πρόβλεψη ενώ η Lasso συρρικνώνει όλους τους συντελεστές προς το 0 με ανάλογα ποσά και κάποιους μέχρι ακριβώς και 0 που οδηγεί σε μοντέλα με λιγότερες μεταβλητές και πιο εύκολη ερμηνεία.

Η Επιλογή του ιδανικού λ είναι ένα δύσκολο πρόβλημα.

Τα ιδανικά/βέλτιστα λ είναι

- «δύσκολο να ρυθμιστούν στην πράξη» (Lederer and Müller, 2015) &
- δεν είναι «πρακτικά εφικτές» (Fan and Tang, 2013).

Fan & Li

- Επιλέγουμε μια μέθοδο κανονικοποίησης όπως Ridge ή Lasso κλπ.
- Χρησιμοποιούμε μια ακολουθία τιμών για το λ και φτιάχνουμε τα αντίστοιχα μοντέλα.
- Μελετάμε τα μοντέλα και επιλέγουμε το κατάλληλο με τα γνωστά κριτήρια όπως το *AIC*, *BIC*.

Η μέθοδος δεν δουλεύει αν οι μεταβλητές *p* είναι πολλές και αυξάνονται εκθετικά με το δείγμα. Τότε, οι Fang and Tang (2013) λένε ότι δεν υπάρχει τρόπος επίλυσης του προβλήματος.

Ο Tibshirani (2013) θεωρεί την δημοφιλή **cross validation** τεχνική ένα εύκολο τρόπο για ευρεση του λ.

cross-validation –διασταυρωμένη επικύρωση:

- επιλέγουμε ένα πλέγμα τιμών του λ,
- υπολογίζουμε το σφάλμα του cross-validation για κάθε τιμή του λ. Επιλέγουμε την τιμη του λ με το μικρότερο σφάλμα.
- Φτιάχνουμε το μοντέλο χρησιμοποιώντας όλες τις παρατηρήσεις και το λ που επιλέχθηκε.

- **ΓΕΝΙΚΑ: Χρήσιμη μέθοδος για σύγκριση μοντέλων.**

Μία εκδοχή: Η μέθοδος παρακράτησης (holdout method), όπου ολόκληρο το dataset χωρίζεται τυχαία σε δύο ξένα μεταξύ τους υποσύνολα: **training & test (πχ ~70%-30% ή ~80%-20%).**

Το σφάλμα πρόβλεψης για το τεστ σετ (με βάση το μοντέλο που προέκυψε από το training set) είναι το ζητούμενο σφάλμα.

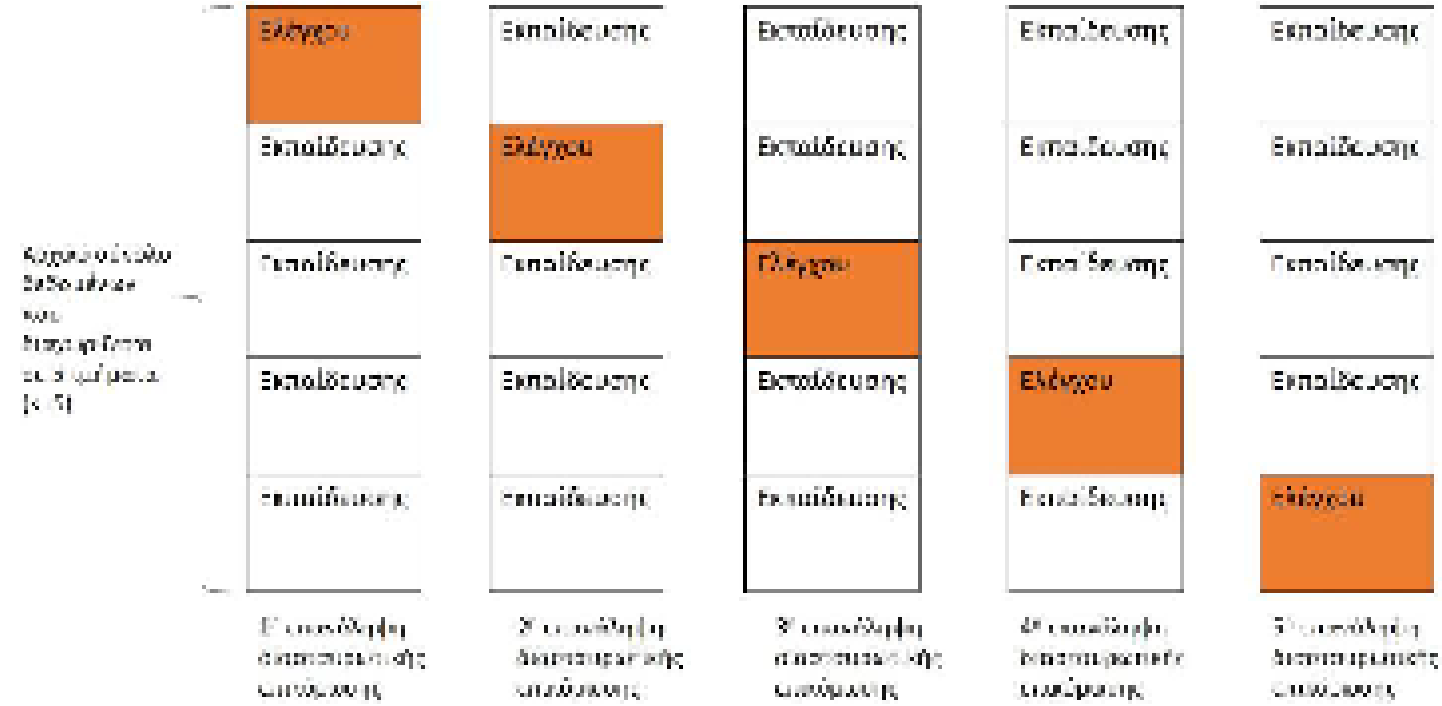
k-fold cross validation: Επαναληπτική διαδικασία όπου το dataset χωρίζεται σε k ξένα υποσύνολα ίσους μεγέθους.

Κάθε ένα από αυτά επιλέγεται διαδοχικά ως **testing set** με τα υπόλοιπα k-1 να χρησιμοποιούνται για **εκπαίδευση**.

Προκύπτουν k μοντέλα και υπολογίζεται το σφάλμα πρόβλεψης με βάση το testing set.

Το τελικό σφάλμα είναι ο μέσος όρος των k σφαλμάτων.

k=2, 5 ή 10.



Εικόνα 0.17 *Διασταυρωμένη επικύρωση 5-πτυχών όπου k=5 που σημαίνει ότι το αρχικό σύνολο δεδομένων θα χωριστεί σε 5 τμήματα. Σε κάθε επανάληψη χρησιμοποιείται διαφορετικό τμήμα δεδομένων για έλεγχο και τα υπόλοιπα για την εκπαίδευση του ίδιου μοντέλου. Μετά από κάθε διαδικασία ελέγχου, θα υπολογιστεί με την επιλεγμένη μετρική ακρίβειας το σφάλμα. Η διαδικασία τερματίζει εάν κάθε ένα από τα 5 τμήματα έχει χρησιμοποιηθεί ως σύνολο ελέγχου.*

ΤΡΟΠΟΙ ΥΠΟΛΟΓΙΣΜΟΥ ΤΟΥ ΣΦΑΛΜΑΤΟΣ

$$MSE = \frac{1}{n}\sum(\hat{y}_i - y_i)^2$$

$$RMSE = \sqrt{\frac{1}{n}\sum(\hat{y}_i - y_i)^2}$$

$$MAE = \frac{1}{n}\sum|\hat{y}_i - y_i|$$

$$MAPE = \frac{1}{n}\sum\left|\frac{\hat{y}_i - y_i}{y_i}\right|$$

ΚΛΑΣΙΚΑ ΚΡΙΤΗΡΙΑ ΕΠΙΛΟΓΗΣ ΜΟΝΤΕΛΩΝ

$$AIC_p = -2\log Lik_p + 2p$$

$$BIC_p = -2\log Lik_p + p\log(n)$$

$$\varphi(p) = -2\log Lik_p + pc\log(\log(n)), \quad c \geq 2$$

Επιλέγεται το μοντέλο (δηλ. το p) που ΕΛΑΧΙΣΤΟΠΟΙΕΙ το κριτηριο

$$R_p^2 = \frac{SSR_p}{SSTO} = 1 - \frac{SSE_p}{SSTO} \quad \& \quad R_{p,adj}^2 = 1 - \frac{SSE_p/(n-p)}{SSTO/(n-1)}$$

Επιλέγεται το μοντέλο (δηλ. το p) που ΜΕΓΙΣΤΟΠΟΙΕΙ το κριτηριο

Η μεγιστοποίηση του adjusted R2 ισοδυναμεί με ΕΛΑΧΙΣΤΟΠΟΙΗΣΗ του MSEp=SSEp/(n-p)