ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΕΦΑΡΜΟΣΜΕΝΩΝ ΜΑΘΗΜΑΤΙΚΩΝ ΚΑΙ ΦΥΣΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

Διπλωματική εργασία

ΠΟΙΝΙΚΟΠΟΙΗΜΕΝΕΣ ΜΕΘΟΔΟΙ ΜΕ ΣΥΝΑΡΤΗΣΕΙΣ ΠΟΙΝΗΣ ΒΑΣΙΣΜΕΝΕΣ ΣΤΗ ΣΥΣΧΕΤΙΣΗ

ΠΑΠΑΓΙΑΝΝΑΚΗΣ ΝΙΚΟΣ

Επιβλέπων: Κουκουβίνος Χρήστος, Καθηγητής ΕΜΠ

Αθήνα, Οκτώβρης 2012

ΠΕΡΙΛΗΨΗ

Τα τελευταία χρόνια έχουν αναπτυχθεί διάφορες μέθοδοι επιλογής μεταβλητών. Σκοπός τους είναι η επιλογή των στατιστικά σημαντικών μεταβλητών, αυτών δηλαδή που επηρεάζουν σημαντικά τη μεταβλητή απόκρισης. Στην παρούσα εργασία παρουσιάζουμε διάφορες ποινικοποιημένες μεθόδους, που αναπτύχθηκαν πρόσφατα, οι οποίες βασίζονται στην εισαγωγή μίας συνάρτησης ποινής στην πιθανοφάνεια. Ειδικότερα, οι συναρτήσεις ποινής που εξετάζουμε, βασίζονται στη συσχέτιση μεταξύ των μεταβλητών.

Στο πρώτο κεφάλαιο, κάνουμε μία εισαγωγή στο γενικό γραμμικό μοντέλο και στις βασικές μεθόδους εκτίμησης των παραμέτρων. Επίσης, παρουσιάζεται το θεωρητικό υπόβαθρο αυτών και αξιολογούμε την απόδοσή τους.

Στο δεύτερο κεφάλαιο, παρουσιάζουμε εκτενώς μία νέα ποινικοποιημένη μέθοδο για γραμμικά μοντέλα, της οποίας η συνάρτηση ποινής βασίζεται στη συσχέτιση μεταξύ των μεταβλητών. Επίσης, συγκρίνουμε την απόδοση της νέας μεθόδου με διάφορες άλλες, εφαρμόζοντας αυτές σε διάφορες προσομοιώσεις και πραγματικά δεδομένα.

Τέλος, στο τρίτο κεφάλαιο, παρουσιάζουμε την επέκταση της νέας μεθόδου στα γενικευμένα γραμμικά μοντέλα. Επίσης, αναφερόμαστε και σε άλλες μεθόδους και αξιολογούμε την αποδοτικότητά τους μέσω προσομοιώσεων και πραγματικών δεδομένων.

ABSTRACT

Over the last years, various methods for variable selection have been developed. The aim is to choose the relevant variables, which have an important influence on the response. In this thesis, we discuss various, recently developed, penalized methods, which are based on a penalty term that is imposed in the likelihood function. In particular, the penalty terms we consider, are based on the correlation between explanatory variables.

In the first chapter, we provide an introduction to the general linear model and to the basic parameter estimation methods. Furthermore, we present the theoretical background of them and we evaluate their performance.

In the second chapter, a novel penalized method with correlation based penalty is presented extensively. We also compare its performance with others methods, through many simulations and real data sets.

At the end, in the third chapter, an extended version of the new method in generalized linear models is presented. We also mention other methods and we evaluate their performance through many simulations and real data sets.

ΕΥΧΑΡΙΣΤΙΕΣ

Η εκπόνηση και ολοκλήρωση της παρούσας διπλωματικής εργασίας δεν θα μπορούσε να πραγματοποιηθεί χωρίς την επίβλεψη και τη συνεισφορά στελεχών του Εθνικού Μετσόβιου Πολυτεχνείου.

Ως εκ τούτου, οφείλω θερμές ευχαριστίες στον Καθηγητή του Ε.Μ.Π. κ. Χρήστο Κουκουβίνο, για την επίβλεψη και καθοδήγηση του, όπως επίσης και για τη δυνατότητα που μου προσέφερε να ασχοληθώ με το επιστημονικό αντικείμενο που με ενδιαφέρει.

Επίσης, θα ήθελα να ευχαριστήσω τον υποψήφιο διδάκτορα Μάνο Ανδρουλάκη, για την πολύτιμη βοήθειά του και το αδιάλειπτο ενδιαφέρον κατά τη διάρκεια εκπόνησης της παρούσας εργασίας.

Τέλος, αυτή η εργασία δεν θα μπορούσε να ολοκληρωθεί χωρίς την υπομονή που έκανε η οικογένειά μου το διάστημα αυτό και τα εφόδια που μου προσέφερε.

ΠΕΡΙΕΧΟΜΕΝΑ

ΚΕΦΑΛΑΙΟ 1

1.1 Εισαγωγή

Έστω ότι $\underset{\sim}{Y}$, μια μεταβλητή που μας ενδιαφέρει και $\underset{\sim}{X}_1, \underset{\sim}{X}_2, ..., \underset{\sim}{X}_p$, ένα σύνολο επεξηγηματικών μεταβλητών ή παραγόντων που αποτελούν διανύσματα n παρατηρήσεων. Το πρόβλημα της επιλογής μεταβλητών εμφανίζεται όταν ο αναλυτής θέλει να μοντελοποιήσει τη σχέση μεταξύ της $\underset{\sim}{Y}$ και ενός υποσυνόλου των $\underset{\sim}{X}_1, \underset{\sim}{X}_2, ..., \underset{\sim}{X}_p$, χωρίς όμως να γνωρίζει ποιο υποσύνολο να επιλέξει. Σκοπός δηλαδή, είναι να επιλεγούν οι παράγοντες που έχουν σημαντική επίδραση στην απόκριση $\underset{\sim}{Y}$. Ιδιαίτερο ενδιαφέρον παρουσιάζει η περίπτωση που συναντάται συχνά στις εφαρμογές, το πλήθος p των υποψηφίων παραγόντων να είναι μεγάλο. Στην πράξη, ο αριθμός των στατιστικά σημαντικών παραγόντων είναι αρκετά μικρότερος σε σχέση με το αρχικό σύνολό τους, μια ιδιότητα γνωστή ως «αρχή της σποραδικότητας των επιδράσεων», (Bickel, 1975). Το πρόβλημα της επιλογής μεταβλητών είναι αρκετά σύνηθες στο πλαίσιο των γραμμικών μοντέλων παλινδρόμησης και των γενικευμένων γραμμικών μοντέλων (Mc Cullagh & Nelder, 1989).

Τα τελευταία χρόνια, έχουν προταθεί αρκετές μέθοδοι και αλγόριθμοι επιλογής μεταβλητών και αποτελούν αναπόσπαστο κομμάτι αρκετών στατιστικών πακέτων. Η χρήση τους γίνεται όλο και περισσότερο αναγκαία, καθότι το μέγεθος των δεδομένων που ανακύπτουν έπειτα από διάφορες μελέτες, συνεχώς μεγαλώνει. Παρότι που ο αριθμός των μεθόδων αυτών είναι αρκετά μεγάλος, το πεδίο της επιλογής μεταβλητών βρίσκεται ακόμα υπό έρευνα και συνεχώς προτείνονται βελτιωμένες ή καινούργιες μέθοδοι.

1.2 Το γενικό γραμμικό μοντέλο παλινδρόμησης

Αρκετές φορές συναντάμε προβλήματα, για τα οποία υπάρχει η υποψία ότι οι τιμές κάποιας μεταβλητής εξαρτώνται από $k \geq 2$ επεξηγηματικές μεταβλητές. Το γενικό γραμμικό μοντέλο, το οποίο περιγράφει αυτή τη σχέση είναι

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + ... + \beta_k x_{ki} + \varepsilon_i = \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij} + \varepsilon_i,\ i = 1, 2, ..., n \quad (1.2.1),$$

οπότε και η απόκριση y_i είναι μια γραμμική συνάρτηση των συντελεστών παλινδρόμησης $_j$, με $j = 1, 2, \ldots, k$. Ως γνωστόν, έχουμε ότι

- y_i, είναι οι τιμές της απόκρισης.
- x_{ij} είναι οι τιμές των επεξηγηματικών μεταβλητών. Υποθέτουμε, όπως και στο απλό γραμμικό μοντέλο, ότι οι μετρήσεις μας δεν υπόκεινται σε σφάλματα.
- $_j$ είναι οι άγνωστες παράμετροι το μοντέλου οι οποίες και πρέπει να εκτιμηθούν.
- £$_i$ είναι τα σφάλματα ή υπόλοιπα, τα οποία αποτελούν τυχαίες μεταβλητές και υποθέτουμε ότι ικανοποιούν τα παρακάτω:
    - $E(£_i) = 0 \quad \forall i$.
    - $Var(£_i) = \pm^2$, δηλαδή τα σφάλματα ικανοποιούν την υπόθεση της ομοιοσκεδαστικότητας.
    - $Cov(£_i, £_j) = 0, \ i \neq j$, δηλαδή τα σφάλματα είναι ασυσχέτιστα.

1.2.1. Εκτίμηση των παραμέτρων του μοντέλου με τη μέθοδο ελαχίστων τετραγώνων

Υποθέτουμε καταρχήν ότι $n > k$. Για να εκτιμηθούν οι παράμετροι $_j$ του μοντέλου, χρησιμοποιείται η μέθοδος ελαχίστων τετραγώνων, (Ανδρουλάκης, 2008). Αυτή η μέθοδος συνίσταται κατά τα γνωστά στην ελαχιστοποίηση του αθροίσματος τετραγώνων των σφαλμάτων

$$S(\ _0, \ _1, \ldots, \ _k) = \sum_{i=1}^{n} £_i^2 = \sum_{i=1}^{n} \left(y_i - \ _0 - \sum_{j=1}^{k} \ _j x_{ij} \right)^2,$$

οπότε τελικά προκύπτουν οι εκτιμητές $\hat{}_0, \hat{}_1, \ldots, \hat{}_k$. Είναι βολικότερο να γράψουμε των εξίσωση (1.2.1) υπό την μορφή πινάκων, ήτοι

$$\underset{\sim}{Y} = \underset{\sim}{X} \ \underset{\sim}{} + \underset{\sim}{£},$$

όπου

$$\underset{\sim}{Y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \qquad \underset{\sim}{X} = \begin{pmatrix} 1 & x_{11} \cdots & x_{1k} \\ 1 & x_{21} \cdots & x_{2k} \\ \vdots & \vdots & \vdots \\ 1 & x_{n1} \cdots & x_{nk} \end{pmatrix}$$

$$\underset{\sim}{} = \begin{bmatrix} 0 \\ 1 \\ \vdots \\ k \end{bmatrix}, \qquad \underset{\sim}{£} = \begin{bmatrix} £_1 \\ £_2 \\ \vdots \\ £_n \end{bmatrix}.$$

Το $\underset{\sim}{Y}$ είναι ένα $n \times 1$ διάνυσμα των παρατηρήσεων, ο $\underset{\sim}{X}$ είναι ένας $n \times k$ πίνακας των επεξηγηματικών μεταβλητών, $\underset{\sim}{}$ είναι ένα $k \times 1$ διάνυσμα των συντελεστών παλινδρόμησης και το $\underset{\sim}{£}$ είναι ένα $n \times 1$ διάνυσμα των τυχαίων σφαλμάτων. Οπότε προκειμένου να βρεθεί η εκτιμήτρια ελαχίστων τετραγώνων $\hat{\underset{\sim}{}}$, πρέπει να ελαχιστοποιηθεί το άθροισμα των τετραγώνων των σφαλμάτων

$$S(\underset{\sim}{}) = \sum_{i=1}^{n} £_i^2 = £'£ = (\underset{\sim}{Y} - \underset{\sim}{X}\,\underset{\sim}{})'(\underset{\sim}{Y} - \underset{\sim}{X}\,\underset{\sim}{}).$$

Τελικά, προκύπτει ότι

$$\hat{\underset{\sim}{}} = (\underset{\sim}{X}'\underset{\sim}{X})^{-1}\underset{\sim}{X}'\underset{\sim}{Y}.$$

Αναφέρουμε και κάποιες ιδιότητες της εκτιμήτριας ελαχίστων τετραγώνων. Καταρχήν

$$E(\hat{\underset{\sim}{}}) = E\left[(\underset{\sim}{X}'\underset{\sim}{X})^{-1}\underset{\sim}{X}'\underset{\sim}{Y}\right]$$

$$= E\left[(\underset{\sim}{X}'\underset{\sim}{X})^{-1}\underset{\sim}{X}'(\underset{\sim}{X}\,\underset{\sim}{} + \underset{\sim}{£})\right]$$

$$= E\left[(\underset{\sim}{X}'\underset{\sim}{X})^{-1}\underset{\sim}{X}'\underset{\sim}{X}\,\underset{\sim}{} + (\underset{\sim}{X}'\underset{\sim}{X})^{-1}\underset{\sim}{X}'\underset{\sim}{£}\right] = \underset{\sim}{},$$

καθότι

$$E(\underset{\sim}{£}) = \underset{\sim}{0}$$

και

$$(\underset{\sim}{X}'\underset{\sim}{X})^{-1}\underset{\sim}{X}'\underset{\sim}{X} = \underset{\sim}{I}.$$

Άρα το $\hat{\underset{\sim}{}}$ αποτελεί αμερόληπτη εκτιμήτρια του $\underset{\sim}{}$. Επίσης έχουμε ότι

$$Var(\hat{\underset{\sim}{}}) = Var\left[(\underset{\sim}{X}'\underset{\sim}{X})^{-1}\underset{\sim}{X}'\underset{\sim}{Y}\right]$$

$$= \left[(\underset{\sim}{X}'\underset{\sim}{X})^{-1}\underset{\sim}{X}'\right] Var(\underset{\sim}{Y}) \left[(\underset{\sim}{X}'\underset{\sim}{X})^{-1}\underset{\sim}{X}'\right]'$$

$$= \pm^2 (\underset{\sim}{X}'\underset{\sim}{X})^{-1}\underset{\sim}{X}'\underset{\sim}{X}(\underset{\sim}{X}'\underset{\sim}{X})^{-1}$$

$$= \pm^2 (\underset{\sim}{X}'\underset{\sim}{X})^{-1}.$$

1.2.2. Εκτίμηση των παραμέτρων του μοντέλου με τη μέθοδο μέγιστης πιθανοφάνειας

Όπως και στην περίπτωση του απλού γραμμικού μοντέλου, αν στις βασικές υποθέσεις προσθέσουμε και ότι τα σφάλματα είναι κανονικά κατανεμημένα, τότε δεν είναι μόνο ασυσχέτιστα αλλά κατ΄ ανάγκη και ανεξάρτητα. Χρησιμοποιώντας διανύσματα, γράφουμε $\underset{\sim}{£} \sim N(\underset{\sim}{0}, \pm^2 \underset{\sim}{I})$, δηλαδή το $\underset{\sim}{£}$ ακολουθεί n-διάστατη πολυμεταβλητή κανονική κατανομή με $E(\underset{\sim}{£}) = \underset{\sim}{0}$ και $Var(\underset{\sim}{£}) = \pm^2 \underset{\sim}{I}$. Σε αυτήν την περίπτωση, η εκτιμήτρια ελαχίστων τετραγώνων $\hat{\underset{\sim}{}}$, ταυτίζεται με την εκτιμήτρια μέγιστης πιθανοφάνειας. Όσον αφορά την τελευταία μέθοδο, ισχύουν τα παρακάτω, (Ανδρουλάκης, 2008).

Η μέθοδος μέγιστης πιθανοφάνειας, προτάθηκε από τον R.A. Fisher (1997). Συγκεκριμένα, έστω ένας πληθυσμός με άγνωστη παράμετρο $\bar{\underset{\sim}{}} = (\bar{}_1, \bar{}_2, ..., \bar{}_k) \in \Theta$ και συνάρτηση πυκνότητας πιθανότητας $f(\underset{\sim}{x} | \bar{\underset{\sim}{}})$. Σκοπός είναι η εκτίμηση της παραμέτρου $\bar{\underset{\sim}{}}$. Οπότε θεωρούμε ένα τυχαίο δείγμα $X_1, X_2, ..., X_n$ από τον πληθυσμό. Αν

$$f(x_1 | \bar{\underset{\sim}{}}), f(x_2 | \bar{\underset{\sim}{}}), ..., f(x_n | \bar{\underset{\sim}{}})$$

είναι η συνάρτηση πυκνότητας πιθανότητας κάθε τιμής του τυχαίου δείγματος, τότε η από κοινού συνάρτηση πυκνότητας πιθανότητας των μεταβλητών $X_1, X_2, ..., X_n$ είναι

$$f(x_1, x_2, ..., x_n | \bar{\underset{\sim}{}}) = f(x_1 | \bar{\underset{\sim}{}}) f(x_2 | \bar{\underset{\sim}{}}) \cdots f(x_n | \bar{\underset{\sim}{}}) \quad (1.2.2.1).$$

Στην περίπτωση συγκεκριμένων παρατηρήσεων $x_1, x_2, ..., x_n$ τυχαίου δείγματος, η (1.2.2.1) είναι συνάρτηση μόνο της παραμέτρου $\underset{\sim}{\theta}$ και συμβολίζεται ως

$$L(\underset{\sim}{\theta} \mid x_1, x_2, ..., x_n) = f(x_1 \mid \underset{\sim}{\theta}) f(x_2 \mid \underset{\sim}{\theta}) \cdots f(x_n \mid \underset{\sim}{\theta}) = \prod_{i=1}^{n} f(x_i \mid \underset{\sim}{\theta}) \quad (1.2.2.2).$$

Η (1.2.2.2) καλείται συνάρτηση πιθανοφάνειας (*likelihood function*) του τυχαίου δείγματος $X_1, X_2, ..., X_n$ και εκφράζει το πόσο «πιθανοφανείς», δηλαδή πόσο σύμφωνες με το συγκεκριμένο δείγμα είναι οι διάφορες τιμές της παραμέτρου $\underset{\sim}{\theta}$.

Η μέθοδος μέγιστης πιθανοφάνειας συνίσταται στην επιλογή της τιμής $\underset{\sim}{\hat{\theta}}$ η οποία μεγιστοποιεί τη συνάρτηση πιθανοφάνειας,

$$L(\underset{\sim}{\hat{\theta}} \mid x_1, x_2, ..., x_n) = \sup_{\underset{\sim}{\theta} \in \Theta} L(\underset{\sim}{\theta} \mid x_1, x_2, ..., x_n).$$

Η τιμή $\underset{\sim}{\hat{\theta}}$ καλείται εκτιμήτρια μέγιστης πιθανοφάνειας της $\underset{\sim}{\theta}$. Μεγιστοποίηση της $L(\underset{\sim}{\theta} \mid x_1, x_2, ..., x_n)$ σημαίνει μεγιστοποίηση της πιθανότητας εμφάνισης των τιμών $x_1, x_2, ..., x_n$ στο δείγμα $X_1, X_2, ..., X_n$. Η τιμή αυτή $\underset{\sim}{\hat{\theta}} = (\hat{\theta}_1, \hat{\theta}_2, ..., \hat{\theta}_k)$, βρίσκεται με λύση των εξισώσεων

$$\frac{\partial \log L(\underset{\sim}{\theta} \mid x_1, x_2, ..., x_n)}{\partial \theta_r} = 0, \quad r = 1, 2, ..., k.$$

Φυσικά, για να είναι η λύση αυτή πράγματι σημείο μεγίστου, θα πρέπει ο Εσσιανός πίνακας

$$\left[\frac{\partial^2 \log L(\underset{\sim}{\theta})}{\partial \theta_i \partial \theta_j} \right]_{k \times k}$$

να είναι γνήσια αρνητικός για $\underset{\sim}{\theta} = \underset{\sim}{\hat{\theta}}$.

1.3 Ποινικοποιημένα ελάχιστα τετράγωνα και ποινικοποιημένη πιθανοφάνεια

1.3.1 Εισαγωγή

Οι πιο γνωστές και συχνότερα χρησιμοποιούμενες μέθοδοι επιλογής μεταβλητών, είναι ως γνωστόν η κατά βήματα απαλοιφή (*stepwise deletion*) και η μέθοδος επιλογής καλύτερου υποσυνόλου (*best subset selection*), (Ανδρουλάκης, 2008). Έχουν όμως το μειονέκτημα ότι αγνοούν τα στοχαστικά σφάλματα που εμφανίζονται κατά τη διαδικασία της επιλογής μεταβλητών καθώς και ότι είναι υπολογιστικά χρονοβόρες. Οι Fan και Li (2001), πρότειναν μια καινούργια μεθοδολογία, βασισμένη στα ποινικοποιημένα ελάχιστα τετράγωνα (*penalized least squares*), η οποία διατηρεί τις καλές ιδιότητες της παλινδρόμησης κορυφογραμμής αλλά και της μεθόδου επιλογής καλύτερου υποσυνόλου. Η μεθοδολογία τους αυτή, επεκτείνεται και σε μοντέλα βασισμένα στη πιθανοφάνεια, όπως π.χ. στην περίπτωση όπου έχουμε δίτιμη απόκριση (*binary response*). Μια γνωστή οικογένεια τέτοιων μοντέλων είναι τα γενικευμένα γραμμικά μοντέλα. Ουσιαστικά τώρα, αυτό που τελικά επιτυγχάνεται, είναι ότι ταυτόχρονα γίνεται και εκτίμηση των παραμέτρων του μοντέλου και μηδενισμός κάποιων, άρα ικανοποιείται ο σκοπός της επιλογής μεταβλητών.

Η διαδικασία της ποινικοποίησης, συνίσταται στην εισαγωγή κάποιων συναρτήσεων ποινής (*penalty functions*), (Ανδρουλάκης, 2008), οι οποίες πρέπει να έχουν τις ακόλουθες ιδιότητες:

- Να είναι ιδιάζουσες (*singular*) στην αρχή ώστε να παράγουν σποραδικές λύσεις (πολλοί εκ των εκτιμηθέντων συντελεστών να έχουν τιμή μηδέν).
- Να ικανοποιούν συγκεκριμένες απαιτήσεις ώστε να παράγουν συνεχή μοντέλα (*continuous models*), οπότε η επιλογή του μοντέλου να χαρακτηρίζεται από σταθερότητα (*stability*).
- Να φράσσονται από μια σταθερά, ώστε να παράγουν σχεδόν αμερόληπτους εκτιμητές για μεγάλους συντελεστές.

Η παλινδρόμηση bridge που προτάθηκε από τους Frank και Friedman (1993), και η μέθοδος LASSO που προτάθηκε από τον Tibshirani (1996) είναι μέλη της μεθόδου των ποινικοποιημένων ελαχίστων τετραγώνων, με τη διαφορά ότι οι σχετικές με τις μεθόδους αυτές, συναρτήσεις ποινής L_q, δεν ικανοποιούν όλες τις προαναφερθείσες απαιτήσεις.

Όπως αναφέραμε και προηγουμένως, η καινούργια μέθοδος επεκτάθηκε και σε μοντέλα βασισμένα στη πιθανοφάνεια (*likelihood-based models*). Η διαφορά σε σχέση με τις παραδοσιακές μεθόδους (όπου συνήθως χρησιμοποιείται τετραγωνική συνάρτηση ποινής), είναι ότι οι νέες συναρτήσεις ποινής είναι συμμετρικές, κυρτές στο $(0,\infty)$ και διακατέχονται από ιδιομορφίες

(*singularities*) στην αρχή. Να σημειωθεί, ότι εν αντιθέσει με τις παραδοσιακές μεθόδους επιλογής μεταβλητών, η νέα μέθοδος έχει ισχυρό θεωρητικό υπόβαθρο. Επίσης, στην εργασία τους, οι Fan και Li (2001), πρότειναν ένα αρκετά αποδοτικό αλγόριθμο βελτιστοποίησης της ποινικοποιημένης πιθανοφάνειας ο οποίος οδηγεί στην εκτίμηση των παραμέτρων και στον υπολογισμό του τυπικού σφάλματος. Δόθηκε μια συγκεκριμένη φόρμουλα υπολογισμού του σφάλματος για τους εκτιμηθέντες συντελεστές χρησιμοποιώντας την μέθοδο *sandwich*. Η μέθοδος αυτή έχει δοκιμαστεί και είναι αρκετά ακριβής για πρακτικούς σκοπούς ακόμα και στη περίπτωση μέτριου μεγέθους δείγματος. Οι προτεινόμενες αυτές διαδικασίες επιλογής συγκρινόμενες με άλλες μεθόδους επιλογής μεταβλητών δίνουν πάντα καλύτερα και ορθότερα αποτελέσματα.

Συνεχίζοντας την περιγραφή των χαρακτηριστικών των μεθόδων αυτών, αναφέρουμε το μεγαλύτερο πλεονέκτημά τους. Συγκεκριμένα, επιλέγουν τις σημαντικές μεταβλητές και εκτιμούν τους συντελεστές τους ταυτόχρονα. Οπότε μπορούν να αναπτυχθούν οι δειγματικές ιδιότητες (*sampling properties*) των μεθόδων, (Ανδρουλάκης, 2008). Στην συνέχεια παρουσιάζουμε πως οι δείκτες σύγκλισης (*rates of convergence*) των προτεινόμενων εκτιμητών της ποινικοποιημένης πιθανοφάνειας (*penalized likelihood estimators*) εξαρτώνται από την παράμετρο κανονικοποίησης. Να σημειωθεί, ότι οι εκτιμητές ποινικοποιημένης πιθανοφάνειας, έχουν τόσο καλή απόδοση όσον αφορά την επιλογή του σωστού μοντέλου, όσο και η διαδικασία προβλεψιμότητας (*oracle procedure*), αρκεί να έχει επιλεγεί σωστά η παράμετρος κανονικοποίησης (*regularization parameter*). Σαν να ήταν δηλαδή γνωστό εξαρχής γνωστό το σωστό υπο-μοντέλο (*submodel*). Αυτό πρακτικά, σημαίνει ότι όταν οι σωστές παράμετροι του μοντέλου έχουν κάποιες μηδενικές συνιστώσες, αυτές εκτιμώνται από τη μέθοδο ως μηδενικές με πιθανότητα να τείνει στη μονάδα. Ενώ όσον αφορά τις μη μηδενικές συνιστώσες, αυτές εκτιμώνται τόσο καλά όπως όταν είναι γνωστό το σωστό υπο-μοντέλο. Αυτό προφανώς αυξάνει την ακρίβεια εκτίμησης τόσο των μηδενικών όσο και των μη μηδενικών συνιστωσών. Οπότε και υπερτερούν της μεθόδου εκτίμησης μέγιστης πιθανοφάνειας. Στη συνέχεια θα γίνει μια εκτενής συζήτηση της όλης μεθοδολογίας.

1.3.2 Επιλογή μεταβλητών μέσω ποινικοποιημένων ελαχίστων τετραγώνων

Θεωρούμε το γνωστό γραμμικό μοντέλο

$$\underset{\sim}{Y} = \underset{\sim}{X}\ \underset{\sim}{} + \underset{\sim}{\text{£}}$$

όπου $\underset{\sim}{Y}$ είναι ένα $n\times 1$ διάνυσμα των παρατηρήσεων, ο $\underset{\sim}{X}$ είναι ένας $n\times d$ πίνακας των επεξηγηματικών μεταβλητών, $\underset{\sim}{\ }$ είναι ένα $d\times 1$ διάνυσμα των συντελεστών παλινδρόμησης και το $\underset{\sim}{\xi}$ είναι ένα $n\times 1$ διάνυσμα των τυχαίων σφαλμάτων. Όπως και στην περίπτωση του μοντέλου γραμμικής παλινδρόμησης, υποθέτουμε ότι τα y_i είναι υπό συνθήκη ανεξάρτητα, δοθέντων των x_{ij}. Επίσης, υποθέτουμε και ότι οι στήλες του πίνακα $\underset{\sim}{X}$ είναι ορθοκανονικές (*orthonormal*). Ο υπολογισμός της εκτιμήτριας γίνεται μέσω της ελαχιστοποίησης της ποσότητας

$$\| \underset{\sim}{Y} - \underset{\sim}{X}\ \underset{\sim}{\ } \|^2,$$

η οποία ισοδυναμεί με την ποσότητα

$$\| \underset{\sim}{\hat{\ }} - \underset{\sim}{\ } \|^2,$$

όπου

$$\underset{\sim}{\hat{\ }} = \underset{\sim}{X}\,'\underset{\sim}{Y}$$

είναι η OLS (*ordinary least squares*) εκτιμήτρια. Θέτοντας τώρα ως

$$\underset{\sim}{z} = \underset{\sim}{X}\,'\underset{\sim}{Y}$$

και έστω ότι

$$\underset{\sim}{\hat{Y}} = \underset{\sim}{X}\underset{\sim}{X}\,'\underset{\sim}{Y},$$

μια μορφή των ποινικοποιημένων ελαχίστων τετραγώνων είναι η εξής:

$$\frac{1}{2}\| \underset{\sim}{Y} - \underset{\sim}{X}\ \underset{\sim}{\ } \|^2 + \text{a}\sum_{j=1}^{d} p_j(|\ \ _j\ |) = \frac{1}{2}\| \underset{\sim}{Y} - \underset{\sim}{\hat{Y}} \|^2 + \frac{1}{2}\sum_{j=1}^{d}(z_j - \ \ _j)^2 + \text{a}\sum_{j=1}^{d} p_j(|\ \ _j\ |) \quad (1.3.2.1).$$

Να σημειωθεί ότι οι συναρτήσεις ποινής p_j στην (1.3.2.1) δεν είναι απαραίτητα οι ίδιες για όλα τα j. Για παράδειγμα μπορεί να θέλουμε να κρατήσουμε ορισμένες σημαντικές μεταβλητές σε ένα παραμετρικό μοντέλο και για αυτό το λόγο να μη θέλουμε να ποινικοποιήσουμε τις αντίστοιχες παραμέτρους τους. Για ευκολία όμως, θεωρούμε ότι οι συναρτήσεις ποινής είναι οι ίδιες για όλους τους συντελεστές, και θα συμβολίζονται ως $p(|\cdot|)$. Επίσης, αντί $\text{a}\,p(|\cdot|)$ θα χρησιμοποιούμε το συμβολισμό $p_{\text{a}}(|\cdot|)$, δείχνοντας έτσι ότι το $p(|\cdot|)$ εξαρτάται από το a.

Το πρόβλημα ελαχιστοποίησης της (1.3.2.1) είναι ισοδύναμο με την ελαχιστοποίηση των συνιστωσών. Οπότε θεωρούμε το παρακάτω πρόβλημα ελαχίστων τετραγώνων

$$\frac{1}{2}(z-\bar{})^2 + p_{\text{a}}(|\bar{}|) \quad (1.3.2.2).$$

Εν συνεχεία, χρησιμοποιώντας τη $Hard$ συνάρτηση ποινής (βλ. σχήμα 1.3.2.1(α))

$$p_{\text{a}}(|\bar{}|) = \text{a}^2 - (|\bar{}| - \text{a})^2 I(|\bar{}| < \text{a}),$$

προκύπτει η $Hard$ εκτιμήτρια (βλ. σχήμα 1.3.2.2α).

$$\hat{\bar{}} = zI(|z| > \text{a}) \quad (1.3.2.3).$$

Με άλλα λόγια, η λύση της (1.3.2.1) είναι

$$z_j I(|z_j| > \text{a})$$

η οποία συμπίπτει με την επιλογή καλύτερου υποσυνόλου και την κατά βήματα πρόσθεση και απαλοιφή στους ορθοκανονικούς σχεδιασμούς. Σημειώνουμε επιπλέον πως η συνάρτηση ποινής $Hard$ είναι ομαλότερη από την συνάρτηση ποινής εντροπίας (*entropy penalty*)

$$p_{\text{a}}(|\bar{}|) = \left(\frac{\text{a}^2}{2}\right) I(|\bar{}| \neq 0),$$

η οποία και αυτή οδηγεί στη λύση (1.3.2.3).

Μια συνάρτηση ποινής για να είναι καλή, πρέπει να δίνει εκτιμητές με τις ακόλουθες ιδιότητες:

- Αμεροληψία: Ο προκύπτων εκτιμητής πρέπει να είναι σχεδόν αμερόληπτος, ιδίως στην περίπτωση όπου η σωστή άγνωστη παράμετρος $_j$ είναι μεγάλη. Αποφεύγεται έτσι η μεροληψία του μοντέλου.
- Σποραδικότητα: Ο προκύπτων εκτιμητής πρέπει να αποτελεί κανόνα περιορισμού (*thresholding rule*), ώστε οι εκτιμηθέντες συντελεστές με μικρή τιμή, να μηδενίζονται. Έτσι, μειώνεται η πολυπλοκότητα του μοντέλου.
- Συνέχεια. Ο προκύπτων εκτιμητής πρέπει είναι συνεχής. Αποφεύγεται κατά αυτόν τον τρόπο η αστάθεια στη πρόβλεψη του μοντέλου.

Ας εξηγήσουμε τώρα τις παραπάνω ιδιότητες. Καταρχήν η πρώτη παράγωγος της (1.3.2.2) ως προς $\bar{}$ είναι

$$\mathrm{sgn}(\bar{\ }) \{ |\bar{\ }| + p'_{\alpha}(|\bar{\ }|) \} - z \, .$$

Παρατηρούμε ότι όταν $p'_{\alpha}(|\bar{\ }|) = 0$ για μεγάλο $|\bar{\ }|$, τότε ο προκύπτων εκτιμητής είναι ίσος με z όταν το $|z|$ είναι επαρκώς μεγάλο. Για αυτό το λόγο, όταν η πραγματική παράμετρος $|\bar{\ }|$ είναι μεγάλη, η τιμή $|z|$ είναι και αυτή μεγάλη και με μεγάλη πιθανότητα. Οπότε, ο PLS (*penalized least squares*) εκτιμητής είναι

$$\hat{\bar{\ }} = z \, ,$$

ο οποίος και είναι σχεδόν αμερόληπτος. Εν συμπεράσματι, η προϋπόθεση $p'_{\alpha}(|\bar{\ }|) = 0$ για μεγάλο $|\bar{\ }|$, είναι μια επαρκής προϋπόθεση για την αμεροληψία μιας μεγάλης πραγματικής παραμέτρου. Όσον αφορά τη δεύτερη ιδιότητα, για να αποτελεί ο προκύπτων εκτιμητής κανόνα περιορισμού, πρέπει να ισχύει ότι

$$\min \{ |\bar{\ }| + p'_{\alpha}(|\bar{\ }|) \} > 0 \, .$$

Το παρακάτω γράφημα 1.3.2.3 παρέχει περισσότερες εξηγήσεις σχετικά με αυτό. Όταν τώρα

$$|z| < \min_{\bar{\ } \neq 0} \{ |\bar{\ }| + p'_{\alpha}(|\bar{\ }|) \}$$

η παράγωγος της (1.3.2.2) είναι θετική για όλα τα θετικά $\bar{\ }$ και αρνητική για όλα τα αρνητικά $\bar{\ }$. Οπότε σε αυτήν την περίπτωση, ο PLS εκτιμητής $\hat{\bar{\ }}$ είναι μηδέν. Όταν όμως $|z| > \min_{\bar{\ } \neq 0} \{ |\bar{\ }| + p'_{\alpha}(|\bar{\ }|) \}$, δύο διασταυρώσεις (*crossings*) μπορούν να υπάρξουν, όπως φαίνεται και στο σχήμα 1.3.2.1. Η μεγαλύτερη είναι ο PLS εκτιμητής. Αυτό συνεπάγεται ότι ικανή και αναγκαία συνθήκη για την ύπαρξη συνέχειας είναι το $\min \{ |\bar{\ }| + p'_{\alpha}(|\bar{\ }|) \}$ να πετυχαίνεται στο μηδέν. Από αυτό αντιλαμβανόμαστε πως η συνάρτηση ποινής που ικανοποιεί τις ιδιότητες της σποραδικότητας και της συνέχειας, πρέπει να είναι ιδιάζουσα (*singular*) στην αρχή.

Είναι γνωστό πως η συνάρτηση ποινής L_2

$$p_{\alpha}(|\bar{\ }|) = \alpha |\bar{\ }|^2$$

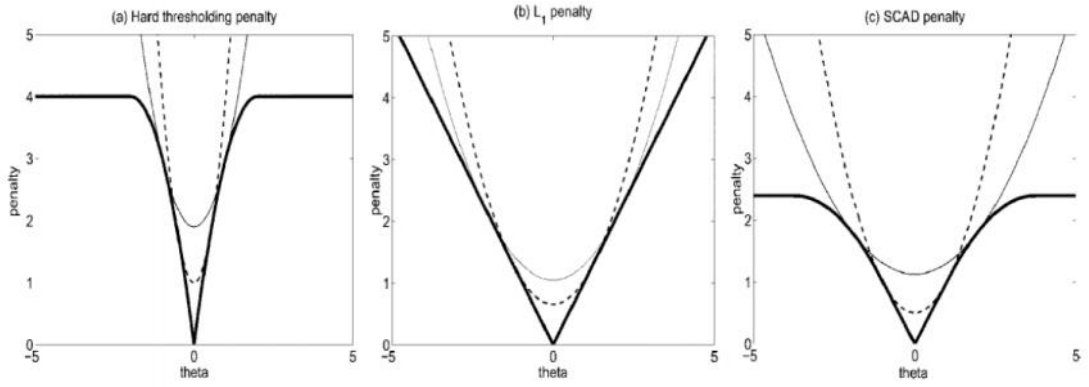
οδηγεί στην παλινδρόμηση κορυφογραμμής. Η συνάρτηση ποινής L_1, οδηγεί στον *soft* οριακό κανόνα

$$\hat{\bar{\ }}_j = \mathrm{sgn}(z_j)(|z_j| - \alpha)_+ \, ,$$

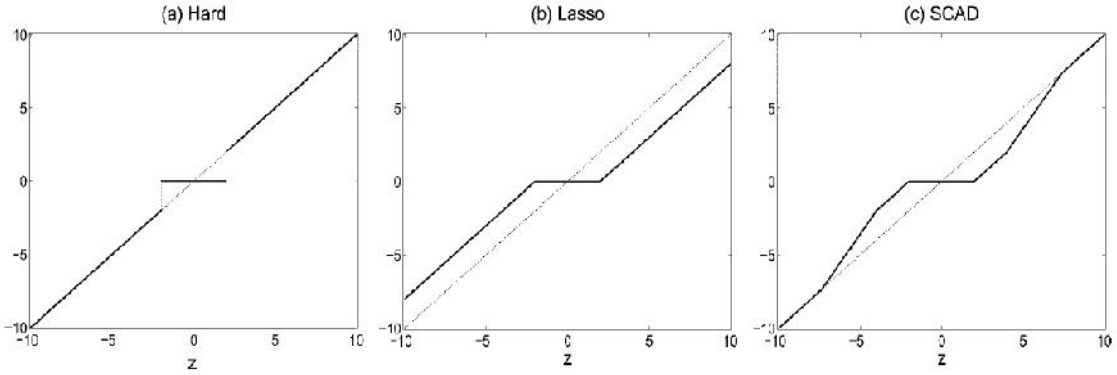
που προτάθηκε από τους Donoho και Johnstone (1994). Η $LASSO$ που προτείνεται από τον Tibshirani (1996, 1997), είναι ο PLS εκτιμητής με συνάρτηση ποινής την L_1. Επίσης, η L_q συνάρτηση ποινής

$$p_a(|\bar{\ }|) = a\,|\bar{\ }|^q$$

οδηγεί στην παλινδρόμηση *bridge* (Frank & Friedman, 1993), (Fu, 1998). Η λύση είναι συνεχής μόνο για $q \geq 1$. Παρόλα αυτά, όταν $q > 1$, δεν παράγεται μια σποραδική λύση (βλ. σχήμα 1.3.2.4(a)). Η μόνη συνεχής λύση με κανόνα περιορισμού σε αυτή την οικογένεια συναρτήσεων είναι με τη συνάρτηση ποινής L_1, αυτό όμως προκύπτει μεταβάλλοντας τον εκτιμητή κατά μια σταθερά λ, άρα χάνεται και η αμεροληψία (βλ. σχήμα 1.3.2.2(b)). Επίσης για $0 \leq q < 1$, δεν ικανοποιείται η συνθήκη της συνέχειας.

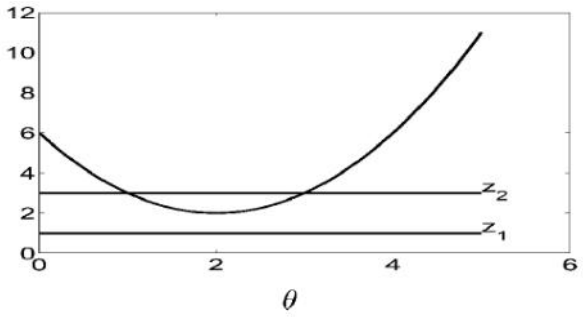


<u>Σχήμα 1.3.2.1</u>: (a) Οι τρεις συναρτήσεις ποινής και οι τετραγωνικές τους προσεγγίσεις.

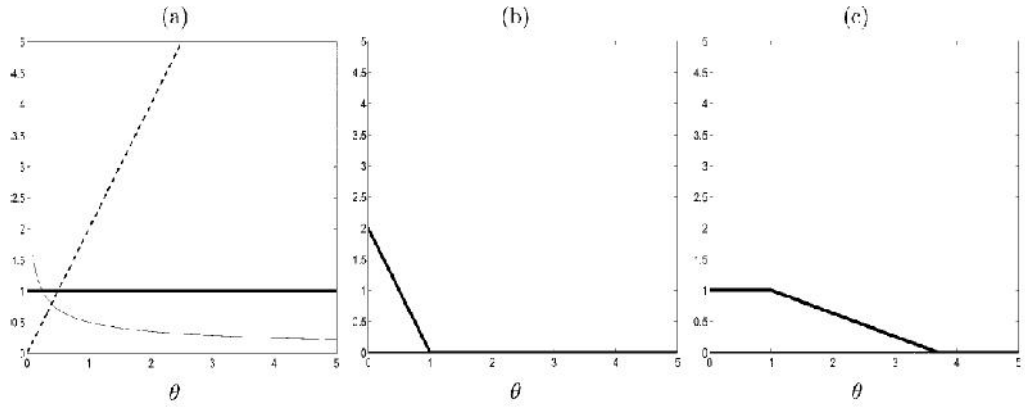


<u>Σχήμα 1.3.2.2</u>: Οι εκτιμήτριες (thresholding functions) (a) Hard, (b) Soft ή LASSO

και (c) Scad, όπου για την τελευταία λ=2 και α=3.7.



***Σχήμα 1.3.2.3**: Η συνάρτηση $^{-} + p_{a}(|^{-}|)$ ως προς $^{-}$.*



***Σχήμα 1.3.2.4**: Οι συναρτήσεις $p'_{a}(|^{-}|)$ ως προς $^{-}$, για (a) τις συναρτήσεις ποινής L_q, (b) τη Hard συνάρτηση ποινής και (c) τη SCAD. Στο (a), η παχιά γραμμή αντιστοιχεί στην L_1, η διακεκομμένη στην $L_{0.5}$ και η λεπτή γραμμή στην L_2 συνάρτηση ποινής.*

1.3.2.1 Η συνάρτηση ποινής SCAD

Οι συναρτήσεις ποινής L_q και Hard δεν ικανοποιούν και τις τρεις απαιτήσεις της αμεροληψίας, της σποραδικότητας και της συνέχειας, (Ανδρουλάκης, 2008). Με σκοπό της βελτίωση της L_1 και της $Hard$, οι Fan και Li (2001) εισήγαγαν μια συνεχής και διαφορίσιμη συνάρτηση ποινής, τη $SCAD$ (*Smoothly Clipped Absolute Deviation penalty*) (βλ. σχήμα 1.3.2.1.(c)), η οποία ορίζεται ως

$$p'_{a}(^{-}) = a\left\{I(^{-} \le a) + \frac{(\sim a - ^{-})_{+}}{(\sim -1)a} I(^{-} > a)\right\}, \text{ για κάποιο } a > 2 \text{ και } ^{-} > 0.$$

Η συγκεκριμένη συνάρτηση δεν ποινικοποιεί υπερβολικά τις μεγάλες τιμές του $\bar{}$ και δίνει μια συνεχής λύση, την

$$\hat{\bar{}} = \begin{cases} \mathrm{sgn}(z)(|z| - \mathrm{a})_+, & |z| \le 2\mathrm{a} \\ \{(\sim -1)z - \mathrm{sgn}(z)\sim \mathrm{a}\}/(\sim -2), & 2\mathrm{a} < |z| \le \sim \mathrm{a} \\ z, & |z| > \sim \mathrm{a} \end{cases} \quad (1.3.2.1.1)$$

Η λύση αυτή δόθηκε από τον Fan (1997), ο οποίος έκανε μια εκτενής συζήτηση για την περίπτωση των κυματοσυναρτήσεων (*wavelets*).

Η λύση (1.3.2.1.1) έχει δύο άγνωστες παραμέτρους, $\sim$ και a. Στην πράξη θα μπορούσαμε να υπολογίσουμε το βέλτιστο ζεύγος $(\sim, \mathrm{a})$ βάσει κάποιων κριτηρίων, όπως της διασταυρωμένης επικύρωσης και της γενικευμένης διασταυρωμένης επικύρωσης. Κάτι που μπορεί να είναι υπολογιστικά χρονοβόρο. Οι Fan και Li (2001), χρησιμοποιώντας εργαλεία Μπεϋζιανής ανάλυσης ρίσκου (*Bayesian risk analysis*), κατέληξαν στην επιλογή του $\sim = 3.7$.

1.3.3 Επιλογή μεταβλητών μέσω ποινικοποιημένης πιθανοφάνειας

Η μέχρι στιγμής αναπτυχθείσα μεθοδολογία, μπορεί να εφαρμοσθεί σε πλήθος στατιστικών μοντέλων, όπως γραμμικά μοντέλα παλινδρόμησης (*linear regression models*), εύρωστα γραμμικά μοντέλα (*robust linear models*) και γενικευμένα γραμμικά μοντέλα βασισμένα στην πιθανοφάνεια (*likelihood-based generalized linear models*). Από και στο εξής, θα θεωρούμε ότι ο πίνακας σχεδιασμού $\underset{\sim}{X} = (x_{ij})$ είναι κανονικοποιημένος, ώστε κάθε στήλη να έχει μέση τιμή 0 και διασπορά 1.

Στο κλασικό μοντέλο παλινδρόμησης οι εκτιμητές ελαχίστων τετραγώνων παράγονται με την ελαχιστοποίηση του αθροίσματος των τετραγώνων των σφαλμάτων. Οπότε η (1.3.2.1) μπορεί να επεκταθεί για την περίπτωση όπου ο πίνακας σχεδιασμού δεν είναι ορθοκανονικός (*orthonormal*). Μια ισοδύναμη μορφή της (1.3.2.1) είναι

$$\frac{1}{2}(\underset{\sim}{Y} - \underset{\sim}{X}\,\underset{\sim}{\ }\,)'(\underset{\sim}{Y} - \underset{\sim}{X}\,\underset{\sim}{\ }\,) + n\sum_{j=1}^{d} p_{\mathrm{a}}(|\ \ _j|) \quad (1.3.3.1).$$

Ελαχιστοποιώντας την (1.3.3.1) ως προς $\underset{\sim}{\ }$, οδηγούμαστε σε έναν εκτιμητή ποινικοποιημένων ελαχίστων τετραγώνων του $\underset{\sim}{\ }$.

Είναι γνωστό τώρα ότι ο OLS εκτιμητής δεν είναι εύρωστος. Μπορούμε όμως να θεωρήσουμε τη συνάρτηση $\cdot$ του Huber (1981), οπότε αντί της ελαχιστοποίησης της (1.3.3.1), μπορούμε να ελαχιστοποιήσουμε την

$$\sum_{i=1}^{n} \cdot (| y_i - \underset{\sim}{x}_i' \; _{\sim} |) + n\sum_{j=1}^{d} p_{\text{a}}(| \quad _j |) \quad \text{(1.3.3.2)},$$

ως προς $_{\sim}$, ώστε να πάρουμε έναν εύρωστο ποινικοποιημένο εκτιμητή του $_{\sim}$.

Στην περίπτωση των γενικευμένων γραμμικών μοντέλων, γίνεται συμπερασματολογία βάσει των εκάστοτε υποβόσκουσων συναρτήσεων πιθανοφάνειας. Με τη βοήθεια τώρα του ποινικοποιημένου εκτιμητή μέγιστης πιθανοφάνειας, μπορούμε να επιλέξουμε σημαντικές μεταβλητές. Έχουμε τα εξής: Καταρχήν, έστω ότι τα δεδομένα $(\underset{\sim}{x}_i, Y_i)$ έχουν συλλεχθεί ανεξάρτητα. Δεδομένων των $\underset{\sim}{x}_i$, η Y_i έχει συνάρτηση πιθανοφάνειας

$$f_i(g(\underset{\sim}{x}_i' \; _{\sim}), y_i),$$

όπου g είναι μια γνωστή συνάρτηση σύνδεσης. Έστω και ότι

$$l_i = \log f_i$$

είναι ο λογάριθμος της πιθανοφάνειας του Y_i. Οπότε μπορούμε να ορίσουμε την ποινικοποιημένη πιθανοφάνεια ως

$$\sum_{i=1}^{n} l_i(g(\underset{\sim}{x}_i' \; _{\sim}), y_i) - n\sum_{j=1}^{d} p_{\text{a}}(| \quad _j |).$$

Η μεγιστοποίηση της ως άνω συνάρτησης, είναι ισοδύναμη με την ελαχιστοποίηση της

$$-\sum_{i=1}^{n} l_i(g(\underset{\sim}{x}_i' \; _{\sim}), y_i) + n\sum_{j=1}^{d} p_{\text{a}}(| \quad _j |) \quad \text{(1.3.3.3)}$$

ως προς $_{\sim}$. Αν αυτό γίνει για κάποια οριακή παράμετρο $^{\text{a}}$, θα πάρουμε τον ποινικοποιημένο εκτιμητή μέγιστης πιθανοφάνειας (*penalized maximum likelihood estimator*).

1.3.3.1 Δειγματοληπτικές και προβλεπτικές ιδιότητες

Σε αυτήν την ενότητα θα αναπτύξουμε την ασυμπτωτική θεωρία του μη κοίλου εκτιμητή ποινικοποιημένης πιθανοφάνειας. Έστω

$$\underset{\sim}{}_0 = (\ _{10},\ldots,\ _{d0})' = (\underset{\sim}{}'_{10}, \underset{\sim}{}'_{20})'.$$

Χωρίς βλάβη της γενικότητας, θεωρούμε ότι

$$\underset{\sim}{}_{20} = \underset{\sim}{0}.$$

Έστω ότι $I(\underset{\sim}{}_0)$ είναι ο πίνακας πληροφορίας του Fisher (*Fisher information matrix*) και έστω $I_1(\underset{\sim}{}_{10}, \underset{\sim}{0})$ η πληροφορία κατά Fisher, γνωρίζοντας ότι $\underset{\sim}{}_{20} = \underset{\sim}{0}$. Αρχικά θα δείξουμε ότι υπάρχει ένας εκτιμητής ποινικοποιημένης πιθανοφάνειας που συγκλίνει στο

$$O_p(n^{-1/2} + \sim_n)$$

όπου

$$\sim_n = \max\left\{ p'_{\mathsf{a}_n}(|\ _{j0}|) :\ \ _{j0} \neq 0 \right\} \text{ (1.3.3.1.1).}$$

Αυτό σημαίνει ότι για τις $Hard$ και $SCAD$ συναρτήσεις ποινής, ο εκτιμητής ποινικοποιημένης πιθανοφάνειας είναι $\sqrt{n}$ -συνεπής (*root-n consistent*) αν $\mathsf{a}_n \to 0$. Επιπλέον θα δείξουμε ότι για τον εκτιμητή αυτόν πρέπει να ισχύει ότι

$$\hat{\underset{\sim}{}}_2 = \underset{\sim}{0}$$

και ότι το $\hat{\underset{\sim}{}}_1$ είναι ασυμπτωτικά της κανονικής κατανομής με πίνακα συνδιασποράς I_1^{-1}, αν

$$n^{1/2}\mathsf{a}_n \to \infty.$$

Αυτό συνεπάγεται ότι ο εκτιμητής ποινικοποιημένης πιθανοφάνειας συμπεριφέρεται τόσο καλά όσο αν ήταν γνωστό ότι $\underset{\sim}{}_{20} = \underset{\sim}{0}$.

Αυτή η προβλεπτική συμπεριφορά του εκτιμητή σχετίζεται άμεσα με το φαινόμενο υπερ-αποδοτικότητας, (*superefficiency phenomenon*). Έστω το απλούστερο γραμμικό μοντέλο παλινδρόμησης

$$\underset{\sim}{Y} = \underset{\sim}{1}_n \ll + \underset{\sim}{£},$$

όπου

$$\underset{\sim}{£} \sim N_n(\underset{\sim}{0}, I_n).$$

Ένας υπερ-αποδοτικός εκτιμητής για το « είναι

$$¢_n = \begin{cases} \bar{Y}, & |\bar{Y}| \geq n^{-1/4} \\ c\bar{Y}, & |\bar{Y}| < n^{-1/4} \end{cases}.$$

Αν θέσουμε το $c = 0$, τότε το $¢_n$ συμπίπτει με τον $Hard$ εκτιμητή με παράμετρο $a_n = n^{-1/4}$. Αυτός ο εκτιμητής υπολογίζει ακριβώς την παράμετρο στο 0 χωρίς να την υπολογίζει σε οποιοδήποτε άλλο σημείο.

Ας γενικεύσουμε τώρα το αποτέλεσμα, θεωρώντας ότι η ποινικοποίηση πραγματοποιείται σε κάθε συνιστώσα του $\underset{\sim}{\ }$. Η περίπτωση όπου κάποιες συνιστώσες δεν ποινικοποιούνται, όπως για παράδειγμα η διασπορά στο γραμμικό μοντέλο, δεν παρουσιάζει κάποιο πρόβλημα. Έστω λοιπόν

$$\underset{\sim}{V}_i = (\underset{\sim}{X}_i, Y_i), \text{ με } i = 1, \ldots, n$$

και ότι $L(\underset{\sim}{\ })$ είναι ο λογάριθμος της πιθανοφάνειας των παρατηρήσεων $\underset{\sim}{V}_1, \ldots, \underset{\sim}{V}_n$. Έστω επίσης ότι

$$Q(\underset{\sim}{\ }) = L(\underset{\sim}{\ }) - n\sum_{j=1}^{d} p_{a_n}(|\ _j|),$$

είναι η ποινικοποιημένη συνάρτηση πιθανοφάνειας. Θα αναφέρουμε στη συνέχεια τα σχετικά θεωρήματα και λήμματα των Fan και Li (2001) των οποίων οι αποδείξεις υπάρχουν στο παράρτημα, αλλά πρωτίστως θα αναφέρουμε κάποιες απαραίτητες υποθέσεις κανονικότητας (*regularity conditions*):

(A) Οι παρατηρήσεις $\underset{\sim}{V}_i$ είναι i.i.d. με συνάρτηση πυκνότητας πιθανότητας $f(\underset{\sim}{V}, \underset{\sim}{\ })$. Η $f(\underset{\sim}{V}, \underset{\sim}{\ })$ έχει μια κοινή βάση και το μοντέλο είναι αναγνωρίσιμο (*identifiable*). Επίσης, η πρώτη και η δεύτερη λογαριθμημένη παράγωγος της f ικανοποιεί τις εξισώσεις

$$E_{\underset{\sim}{\ }}\left[\frac{\partial \log f(\underset{\sim}{V}, \underset{\sim}{\ })}{\partial\ _j}\right] = 0, \text{ για } j = 1, \ldots, d$$

και

$$I_{jk}(\underset{\sim}{\theta}) = E_{\underset{\sim}{\theta}}\left[\frac{\partial}{\partial \theta_j}\log f(\underset{\sim}{V}, \underset{\sim}{\theta})\frac{\partial}{\partial \theta_k}\log f(\underset{\sim}{V}, \underset{\sim}{\theta})\right] = E_{\underset{\sim}{\theta}}\left[-\frac{\partial^2}{\partial \theta_j \partial \theta_k}\log f(\underset{\sim}{V}, \underset{\sim}{\theta})\right].$$

(B) Ο πίνακας πληροφορίας του Fisher

$$I(\underset{\sim}{\theta}) = E\left\{\left[\frac{\partial}{\partial \underset{\sim}{\theta}}\log f(\underset{\sim}{V}, \underset{\sim}{\theta})\right]\left[\frac{\partial}{\partial \underset{\sim}{\theta}}\log f(\underset{\sim}{V}, \underset{\sim}{\theta})\right]'\right\}$$

είναι πεπερασμένος και θετικά ορισμένος στο $\underset{\sim}{\theta} = \underset{\sim}{\theta}_0$.

(C) Υπάρχει ένα ανοικτό υποσύνολο μ του Ω το οποίο περιέχει την πραγματική παράμετρο $\underset{\sim}{\theta}_0$ τέτοιο ώστε για σχεδόν όλα τα $\underset{\sim}{V}$, η συνάρτηση πυκνότητας πιθανότητας $f(\underset{\sim}{V}, \underset{\sim}{\theta})$ επιδέχεται τις παραγώγους τρίτης τάξης

$$\frac{\partial f(\underset{\sim}{V}, \underset{\sim}{\theta})}{\partial \theta_j \partial \theta_k \partial \theta_l}, \text{ για όλα τα } \underset{\sim}{\theta} \in \mu.$$

Επίσης, υπάρχουν συναρτήσεις M_{jkl} τέτοιες ώστε

$$\left|\frac{\partial^3}{\partial \theta_j \partial \theta_k \partial \theta_l}\log f(\underset{\sim}{V}, \underset{\sim}{\theta})\right| \le M_{jkl}(\underset{\sim}{V}), \text{ για όλα τα } \underset{\sim}{\theta} \in \mu,$$

$$\text{όπου } m_{jkl} = E_{\underset{\sim}{\theta}_0}[M_{jkl}] < \infty, \ \forall\, j,k,l.$$

Θεώρημα 1.3.3.1.1

Έστω ότι τα $V_1,...,V_n$ είναι i.i.d. (*independent and identically distributed*), κάθε ένα με συνάρτηση πυκνότητας πιθανότητας $f(V, \theta)$ και ότι ικανοποιούν τις παραπάνω υποθέσεις (A)-(C). Αν

$$\max\left\{| p''_{a_n}(|\quad_{j0}|)|: \quad_{j0} \neq 0\right\} \to 0,$$

τότε υπάρχει ένα τοπικό μέγιστο $\hat{\theta}$ του $Q(\theta)$ τέτοιο ώστε

$$\| \hat{\theta} - \theta_0 \| = O_p(n^{-1/2} + \sim_n),$$

με το $\sim_n$ να δίνεται από την (1.3.3.1.1). Από το θεώρημα αυτό είναι προφανές ότι με μια σωστή επιλογή του a_n θα υπάρξει ένας $\sqrt{n}$ -συνεπής ποινικοποιημένος εκτιμητής. Θα δείξουμε τώρα ότι ο εκτιμητής αυτός έχει την ιδιότητα της σποραδικότητας $\hat{\theta}_2 = 0$.

Λήμμα 1.3.3.1.1

Έστω πάλι ότι τα $V_1,...,V_n$ είναι i.i.d., κάθε ένα με συνάρτηση πυκνότητας πιθανότητας $f(V, \theta)$ και ότι ικανοποιούν τις υποθέσεις (A)-(C). Έστω ότι

$$\liminf_{n\to\infty} \liminf_{\to 0_+} p'_{a_n}(\quad) / a_n > 0 \quad \text{(1.3.3.1.2)}.$$

Αν $a_n \to 0$ και $\sqrt{n} a_n \to \infty$ όσο το $n \to \infty$, τότε με πιθανότητα που τείνει στο 1, για κάθε δοσμένο θ_1 που ικανοποιεί

$$\| \theta_1 - \theta_{10} \| = O_p(n^{-1/2})$$

και για κάθε σταθερά C, ισχύει ότι

$$Q\left\{\begin{pmatrix} \theta_1 \\ 0 \end{pmatrix}\right\} = \max_{\|\theta_2\| \le Cn^{-1/2}} Q\left\{\begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix}\right\}.$$

Ορίζουμε τώρα ως

$$\Sigma = diag\left\{ p''_{a_n}(|\ _{10}|,...,(|\ _{s0}|)\right\}$$

και

$$\underset{\sim}{b} = \left(p'_{a_n}(|\ _{10}|)\,\mathrm{sgn}(\ _{10}),..., p'_{a_n}(|\ _{s0}|)\,\mathrm{sgn}(\ _{s0})\right)'.$$

Θεώρημα 1.3.3.1.2 (Προβλεπτική ιδιότητα)

Θεωρούμε ξανά ότι ότι τα $\underset{\sim}{V}_1,...,\underset{\sim}{V}_n$ είναι i.i.d, κάθε ένα με συνάρτηση πυκνότητας πιθανότητας $f(\underset{\sim}{V},\ _{\sim})$ και ότι ικανοποιούν τις υποθέσεις (A)-(C). Έστω επίσης ότι η συνάρτηση ποινής $p_{a_n}(|\bar{}\ |)$ ικανοποιεί τη συνθήκη (1.6.3.1.2). Αν $a_n \to 0$ και $\sqrt{n}a_n \to \infty$ όσο το $n \to \infty$, τότε με πιθανότητα που τείνει στο 1, οι $\sqrt{n}$ -συνεπείς εκτιμητές $\hat{\ }_{\sim} = \begin{pmatrix} \hat{\ }_{\sim^1} \\ \hat{\ }_{\sim^2} \end{pmatrix}$, του Θεωρήματος 1.3.3.1.1, πρέπει να ικανοποιούν τα παρακάτω:

- Σποραδικότητα (*sparsity*):

$$\hat{\ }_{\sim^2} = \underset{\sim}{0}.$$

- Ασυμπτωτική κανονικότητα (*asymptotic normality*):

$$\sqrt{n}\left(I_1(\ _{\sim 10}) + \Sigma\right)\left\{ \hat{\ }_{\sim^1} - \ _{\sim 10} + \left(I_1(\ _{\sim 10}) + \Sigma\right)^{-1}\underset{\sim}{b}\right\} \to \mathrm{N}\left\{\underset{\sim}{0}, I_1(\ _{\sim 10})\right\},$$

όπου

$$I_1(\ _{\sim 10}) = I_1(\ _{\sim 10}, \underset{\sim}{0})$$

η πληροφορία κατά Fisher, γνωρίζοντας ότι $\ _{\sim^2} = \underset{\sim}{0}$.

Συνεπώς, ο ασυμπτωτικός πίνακας συνδιασποράς του $\hat{\ }_{\sim^1}$ είναι

$$\frac{1}{n}\left\{I_1(\ _{\sim 10}) + \Sigma\right\}^{-1} I_1(\ _{\sim 10})\left\{I_1(\ _{\sim 10}) + \Sigma\right\}^{-1},$$

και για τις συναρτήσεις ποινής που αναπτύχθηκαν στην ενότητα 1.3.2, είναι προσεγγιστικά ίσος με

$$\frac{1}{n} I_1^{-1}(\underset{\sim}{}_{10}) \text{ αν το } \alpha_n \to 0 .$$

Να σημειωθεί ότι για τις $SCAD$ και $Hard$ συναρτήσεις ποινής, αν $\alpha_n \to 0$ τότε $\sim_n = 0$. Οπότε βάσει του Θεωρήματος 1.3.3.1.2, όταν $\sqrt{n}\alpha_n \to \infty$, οι αντίστοιχοι εκτιμητές ποινικοποιημένης πιθανοφάνειας έχουν την προβλεπτική ιδιότητα (*oracle property*) και συμπεριφέρονται τόσο καλά όσο και οι εκτιμητές μέγιστης πιθανοφάνειας, όσον αφορά την εκτίμηση του $\underset{\sim}{}_1$, δεδομένου ότι $\underset{\sim}{}_2 = \underset{\sim}{0}$. Παρόλα αυτά, για την L_1 συνάρτηση ποινής, ισχύει ότι $\sim_n = \alpha_n$. Οπότε, η $\sqrt{n}$-συνέπεια απαιτεί $\alpha_n = O_p(n^{-1/2})$. Όμως, η προβλεπτική ιδιότητα του Θεωρήματος 1.3.3.1.2 απαιτεί $\sqrt{n}\alpha_n \to \infty$. Οι δύο αυτές συνθήκες για τη $LASSO$ δεν ικανοποιούνται ταυτόχρονα. Συνεπώς, δεν ισχύει η προβλεπτική ιδιότητα για την L_1 συνάρτηση ποινής. Αντιθέτως, για την L_q συνάρτηση ποινής, με $q<1$, η προβλεπτική ιδιότητα ισχύει αν έχουμε επιλέξει το σωστό α_n.

Συνεχίζουμε, κάνοντας μια αναφορά περί των συνθηκών κανονικότητας (A)-(C), όσον αφορά τα γενικευμένα γραμμικά μοντέλα. Με μια *canonical link*, η κατανομή του Y δεδομένου ότι $\underset{\sim}{X} = \underset{\sim}{x}$, ανήκει στην *canonical* εκθετική οικογένεια, με συνάρτηση πυκνότητας πιθανότητας

$$f(y, \underset{\sim}{x}, \underset{\sim}{}) = c(y)\exp\left\{\frac{y\underset{\sim}{x}'\underset{\sim}{} - b(\underset{\sim}{x}'\underset{\sim}{})}{\sim(\ddot{})}\right\}.$$

Προφανώς, η συνθήκη (A) ικανοποιείται. Ο πίνακας πληροφορίας του Fisher είναι

$$I(\underset{\sim}{}) = E\left\{b''(\underset{\sim}{x}'\underset{\sim}{})\underset{\sim}{\sim}{x}\underset{\sim}{x}'\right\}/\sim(\ddot{}) .$$

Οπότε αν το $E\left\{b''(\underset{\sim}{x}'\underset{\sim}{})\underset{\sim}{\sim}{x}\underset{\sim}{x}'\right\}$ είναι πεπερασμένο και θετικά ορισμένο, τότε ισχύει και η συνθήκη (B). Επίσης, αν για όλα τα $\underset{\sim}{}$ σε κάποια γειτονιά του $\underset{\sim}{}_0$, ισχύει ότι

$$|b^{(3)}(\underset{\sim}{x}'\underset{\sim}{})| \le M_0(\underset{\sim}{x})$$

για κάποια συνάρτηση $M_0(\underset{\sim}{x})$ που ικανοποιεί

$$E_{\underset{\sim}{}_0}\left\{M_0(\underset{\sim}{x})X_j X_k X_l\right\} < \infty \ \ \forall\, j,k,l ,$$

τότε ισχύει και η συνθήκη (C). Για γενικότερες συναρτήσεις σύνδεσης, παρόμοιες υποθέσεις πρέπει να ικανοποιούνται ώστε να ισχύουν οι συνθήκες (A)-(C). Τα αποτελέσματα των Θεωρημάτων 1.3.3.1.1 και

1.3.3.1.2 μπορούν να προκύψουν και για τις περιπτώσεις των ποινικοποιημένων ελαχίστων τετραγώνων (1.3.3.1) και της ποινικοποιημένης εύρωστης γραμμικής παλινδρόμησης (1.3.3.2).

1.3.3.2 Ο προτεινόμενος αλγόριθμος

Ο Tibshirani (1996) πρότεινε έναν αλγόριθμο για την επίλυση του προβλήματος ελαχίστων τετραγώνων της $LASSO$, ενώ ο Fu (1998) πρότεινε έναν "shooting" αλγόριθμο για την μέθοδο $LASSO$. Στην ενότητα αυτή θα αναπτύξουμε έναν νέο αλγόριθμο που προτάθηκε από τους Fan και Li (2001), με τη βοήθεια του οποίου επιλύονται τα προβλήματα ελαχιστοποίησης (1.3.3.1), (1.3.3.2) και (1.3.3.3). Αυτό γίνεται μέσω τοπικών τετραγωνικών προσεγγίσεων (*local quadratic approximations*). Ο πρώτος όρος των (1.3.3.1), (1.3.3.2) και (1.3.3.3) μπορεί να θεωρηθεί ως μια συνάρτηση απώλειας (*loss function*) του $\underset{\sim}{\ }$. Ας την ονομάσουμε $l(\underset{\sim}{\ })$. Οπότε οι (1.3.3.1), (1.3.3.2) και (1.3.3.3) μπορούν να γραφούν σε μια ενιαία μορφή ως

$$l(\underset{\sim}{\ }) + n\sum_{j=1}^{d} p_{\text{a}}(|\ \ _j|) \quad \text{(1.3.3.2.1).}$$

Οι συναρτήσεις ποινής L_1, $SCAD$ και $Hard$, είναι ιδιάζουσες στην αρχή και δεν έχουν συνεχείς παραγώγους δεύτερης τάξης. Παρόλα αυτά, μπορούν να προσεγγισθούν τοπικά από μια τετραγωνική συνάρτηση ως ακολούθως: Υποθέτουμε ότι έχουμε μια αρχική τιμή $\underset{\sim}{\ }^0$ η οποία είναι πολύ κοντά στην τιμή που ελαχιστοποιεί την (1.3.3.2.1). Αν το $_{j0}$ είναι πολύ κοντά στο 0, τότε θέτουμε $\hat{\ }_j = 0$. Αυτό σημαίνει τη διαγραφή της x_j από το τελικό μοντέλο. Ειδάλλως, χρησιμοποιούμε μια τοπική προσέγγιση της συνάρτησης ποινής $p_{\text{a}}(|\ \ _j|)$, βάσει μιας τετραγωνικής συνάρτησης, ήτοι

$$\left[p_{\text{a}}(|\ \ _j|)\right]' = p'_{\text{a}}(|\ \ _j|)\,\text{sgn}(\ \ _j) \approx \left\{p'_{\text{a}}(|\ \ _{j0}|)/|\ \ _{j0}|\right\}\ \ _j, \text{ όταν } \ \ _j \neq 0.$$

Με άλλα λόγια, έχουμε ότι

$$p_{\text{a}}(|\ \ _j|) \approx p_{\text{a}}(|\ \ _{j0}|) + \frac{1}{2}\left\{p'_{\text{a}}(|\ \ _{j0}|)/|\ \ _{j0}|\right\}\left(\ \ _j^2 - \ \ _{j0}^2\right) \quad \text{(1.3.3.2.2), για } \ \ _j \approx \ \ _{j0}.$$

Το σχήμα 1.3.2.1 της ενότητας 1.3.2 δείχνει τις συναρτήσεις ποινής L_1, $SCAD$ και $Hard$ καθώς και τις προσεγγίσεις τους βάσει της (1.3.3.2.2), για δύο διαφορετικές τιμές του $_{j0}$. Το μόνο μειονέκτημα της προσέγγισης αυτής, είναι ότι από τη στιγμή που κάποιος συντελεστής θα συρρικνωθεί στο 0, θα παραμείνει σε αυτήν την τιμή.

Αν τώρα η $l(\ _{\sim})$ είναι η L_1 συνάρτηση απώλειας, όπως στην (1.3.3.2), τότε δεν έχει συνεχείς μερικές παραγώγους δευτέρας τάξης ως προς $_{\sim}$. Παρόλα αυτά, η ποσότητα $\cdot\ (|\ y-\underset{\sim}{x}'\ _{\sim}\ |)$ στην (1.3.3.2) μπορεί κατά ανάλογο τρόπο να προσεγγισθεί από την

$$\left\{\cdot\ (y-\underset{\sim}{x}'\ _{\sim 0})/(y-\underset{\sim}{x}'\ _{\sim 0})^2\right\}(y-\underset{\sim}{x}'\ _{\sim})^2,$$

αρκεί η αρχική τιμή $_{\sim 0}$ του $_{\sim}$ να είναι αρκετά κοντά στην τιμή ελαχιστοποίησης. Όταν κάποια από τα υπόλοιπα $|\ y-\underset{\sim}{x}'\ _{\sim 0}\ |$ είναι μικρά, η προσέγγιση αυτή δεν είναι καλή. Στην επόμενη ενότητα θα αναλύσουμε κάποιες διαφοροποιήσεις αυτής της προσέγγισης.

Υποθέτουμε στη συνέχεια ότι ο λογάριθμος της πιθανοφάνειας έχει συνεχείς μερικές παραγώγους δευτέρας τάξης ως προς $_{\sim}$. Συνεπώς, είναι εφικτό ο πρώτος όρος της (1.3.3.2.1) να προσεγγισθεί από μια τετραγωνική συνάρτηση. Οπότε, το πρόβλημα ελαχιστοποίησης (1.3.3.2.1) μπορεί να υποβιβασθεί σε ένα τετραγωνικό πρόβλημα ελαχιστοποίησης (*quadratic minimization problem*) και ο αλγόριθμος Newton-Raphson μπορεί να χρησιμοποιηθεί. Πράγματι, η (1.3.3.2.1) προσεγγίζεται (εκτός από έναν σταθερό όρο) από την ποσότητα

$$l(\ _{\sim 0})+\nabla l(\ _{\sim 0})'(\ _{\sim}-\ _{\sim 0})+\frac{1}{2}(\ _{\sim}-\ _{\sim 0})'\nabla^2 l(\ _{\sim 0})(\ _{\sim}-\ _{\sim 0})+\frac{1}{2}n\ _{\sim}'\sum_{a}(\ _{\sim 0})\ _{\sim} \quad \text{(1.3.3.2.3),}$$

όπου

$$\nabla l(\ _{\sim 0})=\frac{\partial l(\ _{\sim 0})}{\partial\ _{\sim}},$$

$$\nabla^2 l(\ _{\sim 0})=\frac{\partial^2 l(\ _{\sim 0})}{\partial\ _{\sim}\partial\ _{\sim}'},$$

$$\sum_{a}(\ _{\sim 0})=diag\left\{p'_{a}(|\ \ _{10}\ |)/|\ \ _{10}\ |,\ldots,p'_{a}(|\ \ _{d0}\ |)/|\ \ _{d0}\ |\right\}.$$

Το τετραγωνικό πρόβλημα ελαχιστοποίησης (1.3.3.2.3), έχει ως λύση την

$$\hat{\underset{\sim}{}}_1 = \hat{\underset{\sim}{}}_0 - \left\{\nabla^2 l(\underset{\sim}{}_0) + n\sum_a (\underset{\sim}{}_0)\right\}^{-1}\left\{\nabla l(\underset{\sim}{}_0) + n\sum_a (\underset{\sim}{}_0)\,\underset{\sim}{}_0\right\}.$$

Όταν επέλθει σύγκλιση του αλγορίθμου, ο εκτιμητής ικανοποιεί τη συνθήκη

$$\frac{\partial l(\hat{\underset{\sim}{}}_0)}{\partial\ _j} + np'_a(|\ \hat{}_{j0}\ |)\,\mathrm{sgn}(\ \hat{}_{j0}) = 0,$$

η οποία αποτελεί την εξίσωση ποινικοποιημένης πιθανοφάνειας, για τα μη μηδενικά στοιχεία του $\hat{\underset{\sim}{}}_0$. Συγκεκριμένα, για το πρόβλημα ποινικοποιημένων ελαχίστων τετραγώνων (1.3.3.1), η λύση βρίσκεται με επαναληπτικό (*iterative*) υπολογισμό της παλινδρόμησης κορυφογραμμής

$$\underset{\sim}{}_1 = \left\{\underset{\sim}{X}'\underset{\sim}{X} + n\sum_a (\underset{\sim}{}_0)\right\}^{-1}\underset{\sim}{X}'\underset{\sim}{Y}.$$

Ομοίως, η λύση της (1.3.3.2) προκύπτει με επαναληπτικό υπολογισμό της

$$\underset{\sim}{}_1 = \left\{\underset{\sim}{X}'\underset{\sim}{W}\underset{\sim}{X} + \frac{1}{2}n\sum_a (\underset{\sim}{}_0)\right\}^{-1}\underset{\sim}{X}'\underset{\sim}{W}\underset{\sim}{Y},$$

όπου

$$\underset{\sim}{W} = diag\left\{\cdot\ (|\ y_1 - \underset{\sim}{x}_1'\ \underset{\sim}{}_0\ |)/(y_1 - \underset{\sim}{x}_1'\ \underset{\sim}{}_0)^2, \ldots, \cdot\ (|\ y_n - \underset{\sim}{x}_n'\ \underset{\sim}{}_0\ |)/(y_n - \underset{\sim}{x}_n'\ \underset{\sim}{}_0)^2\right\}.$$

Όπως και στην περίπτωση του εκτιμητή μέγιστης πιθανοφάνειας, έχοντας μια καλή αρχική τιμή $\underset{\sim}{}_0$, η μονοβηματική διαδικασία μπορεί να είναι εξίσου αποδοτική όσο και η πλήρως επαναληπτική διαδικασία όπου παίρνουμε τον εκτιμητή ποινικοποιημένης πιθανοφάνειας, κάνοντας χρήση του αλγορίθμου Newton-Raphson. Αν τώρα θεωρήσουμε ως $\underset{\sim}{}^{(k-1)}$ μια καλή αρχική τιμή στο k βήμα, ο επόμενος επαναληπτικός υπολογισμός μπορεί να θεωρηθεί ως μονοβηματική διαδικασία, άρα ο προκύπτων εκτιμητής εξακολουθεί να μπορεί να είναι το ίδιο αποδοτικός όσο αυτός που θα προέκυπτε με την πλήρως επαναληπτική μέθοδο. Συμπερασματικά, ο εκτιμητής που θα προκύψει με τον αλγόριθμο που αναφέραμε κάνοντας λίγες επαναλήψεις, μπορεί να θεωρηθεί ως εκτιμητής ενός βήματος και θα έχει την ίδια απόδοση. Οπότε βάσει αυτού του σκεπτικού, δεν χρειάζεται να επαναλάβουμε τον αλγόριθμο μέχρι να επέλθει σύγκλιση, αρκεί οι αρχικές εκτιμήσεις να είναι καλές. Ως αρχικές εκτιμήσεις τώρα, μπορούν να δοθούν αυτές του πλήρους μοντέλου, αρκεί να μην είναι υπερβολικά παραμετροποιημένες.

1.3.3.3 Υπολογισμός του τυπικού σφάλματος

Τα τυπικά σφάλματα των εκτιμηθέντων παραμέτρων μπορούν άμεσα να υπολογισθούν, λόγω του ότι γίνεται ταυτόχρονη εκτίμηση παραμέτρων και επιλογή μεταβλητών. Ο *sandwich* τύπος μπορεί να χρησιμοποιηθεί για την εκτίμηση της συνδιασποράς του $\hat{}_{1}$, η μη εξαφανισμένη συνιστώσα του $\hat{}$. Οπότε έχουμε,

$$\widehat{\mathrm{cov}}(\hat{}_{1}) = \left\{\nabla^2 l(\hat{}_{1}) + n\sum\nolimits_{a}(\hat{}_{1})\right\}^{-1} \widehat{\mathrm{cov}}\left\{\nabla l(\hat{}_{1})\right\}\left\{\nabla^2 l(\hat{}_{1}) + n\sum\nolimits_{a}(\hat{}_{1})\right\}^{-1} \quad (1.3.3.3.1).$$

Ο τύπος αυτός είναι αρκετά ακριβής και για μέτρια μεγέθη δειγμάτων.

Όταν χρησιμοποιείται η L_1 συνάρτηση απώλειας στην εύρωστη παλινδρόμηση, πρέπει να πραγματοποιηθούν κάποιες τροποποιήσεις στον αλγόριθμο καθώς επίσης και στον αντίστοιχο *sandwich* τύπο. Στην περίπτωση όπου $\cdot\,(x) = |x|$, τα διαγώνια στοιχεία του W είναι

$$\{|r_i|^{-1}\}, \text{ με } r_i = y_i - x_i'\,{}_{0} \text{ και } i = 1,\ldots,n.$$

Οπότε για μια δοθείσα τιμή του ${}_{0}$, όταν κάποια από τα υπόλοιπα $\{r_i\}$ είναι κοντά στο 0, αυτά τα σημεία αποκτούν πολύ βάρος. Για αυτό το λόγο αντικαθίσταται το βάρος με

$$(\sim_n + |r_i|^{-1}).$$

Στις εφαρμογές που έκαναν οι Fan και Li, χρησιμοποίησαν ως $\sim_n$ το $2n^{-1/2}$ *quantile* των απολύτων τιμών των υπολοίπων, $\{|r_i|\}$. Οπότε το $\sim_n$ άλλαζε σε κάθε επανάληψη.

1.3.3.4 Έλεγχος τη σύγκλισης του αλγορίθμου

Οι Fan και Li, απέδειξαν με χρήση του προγράμματος MATLAB ότι όντως ο αλγόριθμος που πρότειναν συγκλίνει στη σωστή λύση. Συγκεκριμένα, χρησιμοποίησαν ένα διάνυσμα $$ διάστασης 100, αποτελούμενο από 50 μηδενικά και 50 μη μηδενικά στοιχεία που και δημιουργήθηκαν από την κατανομή $N(0,5^2)$. Επίσης χρησιμοποίησαν έναν 100×100 ορθοκανονικό πίνακα σχεδιασμού, για το λόγο ότι τα ποινικοποιημένα ελάχιστα τετράγωνα (PLS) έχουν τότε μαθηματική λύση κλειστής μορφής, οπότε και ήταν εφικτή η σύγκρισή της με αυτήν της αλγοριθμικής μεθόδου τους. Το διάνυσμα

των αποκρίσεων $\underset{\sim}{Y}$ δημιουργήθηκε βάσει του γραμμικού μοντέλου $\underset{\sim}{Y} = \underset{\sim}{X}\ \underset{\sim}{} + \underset{\sim}{\varepsilon}$. Τα αποτελέσματα ήταν τα εξής: Το MATLAB χρειάστηκε 0.27, 0.39 και 0.16 sec για να επέλθει σύγκλιση όσον αφορά τα PLS με τη $SCAD$, L_1 και $Hard$ συνάρτηση ποινής αντίστοιχα. Επίσης, ο αριθμός των επαναλήψεων ήταν 30, 30 και 5 αντίστοιχα. Να σημειωθεί, ότι στη δέκατη επανάληψη, ο PLS εκτιμητής ήταν ήδη αρκετά κοντά στη σωστή τιμή.

1.3.4 Αριθμητικές συγκρίσεις

Στην ενότητα αυτή, θα συγκρίνουμε την απόδοση των προτεινόμενων μεθόδων με τις ήδη υπάρχουσες και θα ελέγξουμε την ακρίβεια της μεθόδου εύρεσης του τυπικού σφάλματος. Επίσης θα αναφέρουμε και κάποιες μελέτες προσομοίωσης (*simulation studies*) που έκαναν οι Fan και Li χρησιμοποιώντας τις ποινικοποιημένες μεθόδους.

1.3.4.1 Σφάλμα πρόβλεψης και σφάλμα μοντέλου

Το σφάλμα πρόβλεψης (*prediction error*) ορίζεται ως το μέσο σφάλμα στην πρόβλεψη του Y, δεδομένου νέου $\underset{\sim}{x}$ (που προφανώς δεν χρησιμοποιήθηκε στην κατασκευή της εξίσωσης πρόβλεψης). Υπάρχουν δύο περιπτώσεις, το X να είναι τυχαίο (*random*) και το X να είναι ελεγχόμενο (*controlled*). Στην πρώτη περίπτωση, τόσο το Y όσο και το $\underset{\sim}{x}$ είναι τυχαία επιλεγμένα. Στην δεύτερη περίπτωση, ο πίνακας σχεδιασμού επιλέγεται από τους πειραματιστές και μόνο το Y είναι τυχαίο. Στο εξής θα θεωρούμε ότι το X είναι τυχαίο.

Σε αυτήν την περίπτωση, τα δεδομένα $(\underset{\sim}{x}_i, Y_i)$ θεωρούνται τυχαίο δείγμα από κάποια κατανομή. Τότε, αν $\hat{«}(\underset{\sim}{x})$ είναι η πρόβλεψη βάσει των δεδομένων που έχουμε στην κατοχή μας, το σφάλμα πρόβλεψης ορίζεται ως

$$PE(\hat{«}) = E\{Y - \hat{«}(\underset{\sim}{x})\}^2.$$

Ο παραπάνω τύπος μπορεί να αναλυθεί ως

$$PE(\hat{«}) = E\{Y - E(Y \mid \underset{\sim}{x})\}^2 + E\{E(Y \mid \underset{\sim}{x}) - \hat{«}(\underset{\sim}{x})\}^2.$$

Ο πρώτος όρος είναι το σφάλμα πρόβλεψης λόγω του θορύβου στα δεδομένα και ο δεύτερος λόγω της έλλειψης προσαρμογής (*lack of fit*) του μοντέλου. Αυτός ο δεύτερος όρος ονομάζεται σφάλμα μοντέλου (*model error*) και συμβολίζεται ως $ME(\hat{\alpha})$. Να σημειώσουμε ότι αν $Y = \underset{\sim}{x}' \underset{\sim}{} + e$, με $E(e \mid \underset{\sim}{x}) = 0$, τότε

$$ME(\hat{\alpha}) = (\hat{\underset{\sim}{}} - \underset{\sim}{})' E(\underset{\sim}{x}\underset{\sim}{x}')(\hat{\underset{\sim}{}} - \underset{\sim}{}) .$$

1.3.4.2 Επιλογή των οριακών παραμέτρων

Οι Fan και Li, προκειμένου να εκτιμήσουν τη ρυθμιστική (*tuning*) παράμετρο $\bar{\underset{\sim}{}}$, όπου $\bar{\underset{\sim}{}} = (\text{a}, \sim)$ για τη $SCAD$ συνάρτηση ποινής και τη $\bar{\underset{\sim}{}} = \text{a}$ για τη $LASSO$ και $Hard$, χρησιμοποίησαν δύο μεθόδους. Την πενταπλή (*fivefold*) διασταυρωμένη επικύρωση και τη γενικευμένη διασταυρωμένη επικύρωση. Θα αναπτύξουμε τις δύο αυτές διαδικασίες για την περίπτωση των γραμμικών μοντέλων παλινδρόμησης. Η επέκταση των διαδικασιών αυτών σε εύρωστα γραμμικά μοντέλα παλινδρόμησης καθώς και γραμμικά μοντέλα βασισμένα στην πιθανοφάνεια, δεν εμπεριέχει ιδιαίτερες δυσκολίες.

Στη μέθοδο της πενταπλής διασταυρωμένης επικύρωσης, συμβολίζουμε ως T το σύνολο των δεδομένων και ως $T - T^{\neg}$ και $T^{\neg}$ το σύνολο εκπαίδευσης (*training set*) και το σύνολο ελέγχου (*test set*) αντίστοιχα, με $\neg = 1, \ldots, 5$. Για κάθε $\bar{\underset{\sim}{}}$ και $\neg$, βρίσκουμε τον εκτιμητή $\hat{\underset{\sim}{}}^{(\neg)}(\bar{\underset{\sim}{}})$ του $\underset{\sim}{}$, χρησιμοποιώντας το σύνολο εκπαίδευσης $T - T^{\neg}$. Εν συνεχεία, εφαρμόζουμε το κριτήριο της διασταυρωμένης επικύρωσης

$$CV(\bar{\underset{\sim}{}}) = \sum_{\neg=1}^{5} \sum_{(y_k, \underset{\sim}{x}_k) \in T^{\neg}} \left\{ y_k - \underset{\sim}{x}_k' \, \hat{\underset{\sim}{}}^{(\neg)}(\bar{\underset{\sim}{}}) \right\}^2$$

και βρίσκουμε το $\hat{\bar{\underset{\sim}{}}}$ που ελαχιστοποιεί το $CV(\bar{\underset{\sim}{}})$.

Στη μέθοδο της γενικευμένης διασταυρωμένης επικύρωσης, μετατρέπουμε τη λύση ως

$$\underset{\sim}{}^{1}(\bar{\underset{\sim}{}}) = \left\{ \underset{\sim}{X}' \underset{\sim}{X} + n \sum\nolimits_{\text{a}} (\underset{\sim}{}^{0}) \right\}^{-1} \underset{\sim}{X}' \underset{\sim}{Y} .$$

Οπότε η προσαρμοσμένη τιμή $\hat{\underset{\sim}{Y}}$ του $\underset{\sim}{Y}$ είναι

$$\underset{\sim}{X} \left\{ \underset{\sim}{X}' \underset{\sim}{X} + n \sum\nolimits_{\text{a}} (\underset{\sim}{}^{0}) \right\}^{-1} \underset{\sim}{X}' \hat{\underset{\sim}{Y}}$$

και μπορούμε να θεωρήσουμε ως πίνακα προβολής τον

$$\underset{\sim}{P}_{\underset{\sim}{X}}\{\hat{\underset{\sim}{}}(\bar{\underset{\sim}{}})\} = \underset{\sim}{X}\left\{\underset{\sim}{X}'\underset{\sim}{X} + n\sum\nolimits_{a}(\hat{\underset{\sim}{}})\right\}^{-1}\underset{\sim}{X}'.$$

Ορίζοντας τώρα το πλήθος των σημαντικών παραμέτρων στην προσαρμογή του ποινικοποιημένου μοντέλου ελαχίστων τετραγώνων ως

$$e(\bar{\underset{\sim}{}}) = tr[\underset{\sim}{P}_{\underset{\sim}{X}}\{\hat{\underset{\sim}{}}(\bar{\underset{\sim}{}})\}],$$

το κριτήριο της γενικευμένης διασταυρωμένης επικύρωσης είναι

$$GCV(\bar{\underset{\sim}{}}) = \frac{1}{n}\frac{\|\underset{\sim}{Y} - \underset{\sim}{X}\,\hat{\underset{\sim}{}}(\bar{\underset{\sim}{}})\|^2}{\{1 - e(\bar{\underset{\sim}{}})/n\}^2}$$

και

$$\hat{\bar{\underset{\sim}{}}} = \arg\min\{GCV(\bar{\underset{\sim}{}})\}.$$

1.3.4.3 Προσομοιώσεις

Οι Fan και Li, στα ακόλουθα παραδείγματα προσομοιώσεων, σύγκριναν τις προτεινόμενες μεθόδους επιλογής μεταβλητών με τις ακόλουθες μεθόδους:

Α) Ελάχιστα τετράγωνα.

Β) Παλινδρόμηση κορυφογραμμής.

Γ) Επιλογή καλύτερου υποσυνόλου.

Δ) *Garrote*.

Οι προσομοιώσεις έγιναν με χρήση του MATLAB. Χρησιμοποιήθηκε επίσης η γενικευμένη διασταυρωμένη επικύρωση για την εκτίμηση των οριακών παραμέτρων.

Προσομοίωση 1-Γραμμική παλινδρόμηση: Δημιουργήθηκαν 100 σύνολα δεδομένων, αποτελούμενα από n παρατηρήσεις, βάσει του μοντέλου

$$\underset{\sim}{Y} = \underset{\sim}{x}' \underset{\sim}{} + \pm\underset{\sim}{\pounds},$$

όπου τα $\underset{\sim}{x}$ και $\underset{\sim}{\pounds}$ είναι της Τυποποιημένης Κανονικής κατανομής και $\underset{\sim}{} = (3,1.5,0,0,2,0,0,0)'$. Η συσχέτιση μεταξύ των x_i και x_j είναι $\circ^{|i-j|}$ με $\circ = 0.5$. Αρχικά, έγινε η επιλογή του $n = 40$ και του $\pm = 3$. Έπειτα, μειώθηκε το $\pm$ σε 1 και το n αυξήθηκε στις 60 παρατηρήσεις. Το σφάλμα του μοντέλου συγκρίθηκε με αυτό του εκτιμητή ελαχίστων τετραγώνων. Η διάμεσος των σχετικών σφαλμάτων του μοντέλου (*Median of Relative Model Errors* $-MRME$) από 100 προσομοιωμένα σύνολα δεδομένων, υπάρχει στον πίνακα 1.3.4.3.1. Επίσης, στον ίδιο πίνακα φαίνεται και ο μέσος αριθμός των μηδενικών συντελεστών, με τη στήλη «correct» να αντιστοιχεί στο μέσο αριθμό των σωστά εκτιμηθέντων ως μηδενικοί συντελεστών, ενώ η στήλη «incorrect» αντιστοιχεί σε αυτούς που λανθασμένα εκτιμήθηκαν ως μηδενικοί.

		Avg. No. of 0 Coefficients	
Method	*MRME (%)*	*Correct*	*Incorrect*
$n = 40, \sigma = 3$			
SCAD[1]	72.90	4.20	.21
SCAD[2]	69.03	4.31	.27
LASSO	63.19	3.53	.07
Hard	73.82	4.09	.19
Ridge	83.28	0	0
Best subset	68.26	4.50	.35
Garrote	76.90	2.80	.09
Oracle	33.31	5	0
$n = 40, \sigma = 1$			
SCAD[1]	54.81	4.29	0
SCAD[2]	47.25	4.34	0
LASSO	63.19	3.51	0
Hard	69.72	3.93	0
Ridge	95.21	0	0
Best subset	53.60	4.54	0
Garrote	56.55	3.35	0
Oracle	33.31	5	0
$n = 60, \sigma = 1$			
SCAD[1]	47.54	4.37	0
SCAD[2]	43.79	4.42	0
LASSO	65.22	3.56	0
Hard	71.11	4.02	0
Ridge	97.36	0	0
Best subset	46.11	4.73	0
Garrote	55.90	3.38	0
Oracle	29.82	5	0

***Πίνακας 1.3.4.3.1**: Αποτελέσματα προσομοιώσεων για το γραμμικό μοντέλο παλινδρόμησης. Για τη SCAD[1] το α επιλέχθηκε βάσει της GCV και για τη SCAD[2] έχει την τιμή 3.7.*

Από τον παραπάνω πίνακα, παρατηρούμε ότι όταν ο θόρυβος είναι υψηλός και το μέγεθος του δείγματος μικρό, η $LASSO$ έχει την καλύτερη απόδοση. Επίσης μειώνει σημαντικά τόσο το σφάλμα του μοντέλου όσο και την πολυπλοκότητά του. Αυτό ισχύει και για τις υπόλοιπες μεθόδους επιλογής μεταβλητών, ενώ αντιθέτως, η παλινδρόμηση κορυφογραμμής μειώνει μόνο το σφάλμα του μοντέλου. Όταν όμως μειώθηκε ο θόρυβος, η $SCAD$ είναι αποδοτικότερη από τη $LASSO$ και τη $Hard$. Η παλινδρόμηση κορυφογραμμής έχει κακή απόδοση ενώ η μέθοδος επιλογής καλύτερου υποσυνόλου έχει παρόμοια απόδοση με τη $SCAD$. Επίσης, η *garrote* έχει γενικά καλή απόδοση. Να σημειώσουμε και ότι η *SCAD* είχε πολύ καλά αποτελέσματα με επιλογή του $\sim = 3.7$ (βλ. αποτελέσματα για $SCAD^1$ και $SCAD^2$), η οποία τιμή χρησιμοποιήθηκε και στις επόμενες προσομοιώσεις. Τελειώνοντας, συμπεραίνουμε ότι αναμένεται η *SCAD* να έχει τόσο καλά αποτελέσματα όσο αυτά του *oracle* εκτιμητή (ο οποίος επίσης χρησιμοποιήθηκε ώστε να συγκριθεί με τις προτεινόμενες μεθόδους), καθώς το μέγεθος του δείγματος αυξάνει.

Όσον αφορά τώρα την ακρίβεια της μεθόδου υπολογισμού του τυπικού σφάλματος (1.3.3.3.1), έχουμε τα εξής: Η διάμεσος των απολύτων τιμών της απόκλισης των 100 εκτιμηθέντων συντελεστών των 100 συνόλων δεδομένων, διαιρεμένη με 0.6745, συμβολιζόμενη ως *SD*, βρίσκεται στον πίνακα 1.8.1.2 και μπορεί να θεωρηθεί ως το πραγματικό τυπικό σφάλμα. Η διάμεσος των 100 αυτών εκτιμηθέντων *SDs*, συμβολίζεται με SD_m και η διάμεσος των απολύτων τιμών του σφάλματος της απόκλισης των 100 εκτιμημένων τυπικών σφαλμάτων διαιρεμένη με 0.6745, συμβολίζεται με SD_{mad} αποτελούν μια αποτίμηση της συνολικής απόδοσης της (1.3.3.3.1). Ο πίνακας 1.3.4.3.2 περιέχει τα αποτελέσματα για τους μη μηδενικούς συντελεστές, στην περίπτωση όπου $n = 60$. Στην περίπτωση όπου $n = 40$, είχαμε παρόμοια αποτελέσματα. Βάσει του πίνακα αυτού, συμπεραίνουμε ότι ο *sandwich* τύπος 1.3.4.3.2 είναι αρκετά αποτελεσματικός.

	$\hat{\beta}_1$		$\hat{\beta}_2$		$\hat{\beta}_5$	
Method	*SD*	SD_m (SD_{mad})	*SD*	SD_m (SD_{mad})	*SD*	SD_m (SD_{mad})
$SCAD^1$	.166	.161 (.021)	.170	.160 (.024)	.148	.145 (.022)
$SCAD^2$	.161	.161 (.021)	.164	.161 (.024)	.151	.143 (.023)
LASSO	.164	.154 (.019)	.173	.150 (.022)	.153	.142 (.021)
Hard	.169	.161 (.022)	.174	.162 (.025)	.178	.148 (.021)
Best subset	.163	.155 (.020)	.152	.154 (.026)	.152	.139 (.020)
Oracle	.155	.154 (.020)	.147	.153 (.024)	.146	.137 (.019)

<u>Πίνακας 1.3.4.3.2</u>: Τυπικές αποκλίσεις των εκτιμητών στο γραμμικό μοντέλο παλινδρόμησης (n=60).

Προσομοίωση 2-Εύρωστη γραμμική παλινδρόμηση: Δημιουργήθηκαν 100 σύνολα δεδομένων αποτελούμενα από 60 παρατηρήσεις, βάσει του μοντέλου

$$Y = \underset{\sim}{x}' \underset{\sim}{} + £ ,$$

με τα ίδια $\underset{\sim}{}$ και $\underset{\sim}{x}$ όπως και στην προηγούμενη προσομοίωση. Το $£$ είναι της Τυποποιημένης Κανονικής κατανομής με ένα ποσοστό 10% άτυπων σημείων (*outliers*) της κατανομής *Cauchy*. Τα αποτελέσματα βρίσκονται στον πίνακα 1.3.4.3.3. Βλέπουμε ότι την καλύτερη απόδοση την έχει η *SCAD*. Επίσης, οι αληθείς και οι εκτιμώμενες βάσει της (1.3.3.3.1) τυπικές αποκλίσεις των εκτιμητών βρίσκονται στον πίνακα 1.3.4.3.4, όπου και καταδεικνύεται η πολύ καλή απόδοση της μεθόδου.

		Avg. No. of 0 Coefficients	
Method	*MRME (%)*	*Correct*	*Incorrect*
SCAD ($a = 3.7$)	35.52	4.71	0
LASSO	52.80	4.29	0
Hard	47.22	4.70	0
Best subset	41.53	4.85	.18
Oracle	23.33	5	0

***Πίνακας 1.3.4.3.3**: Αποτελέσματα προσομοίωσης για το εύρωστο γραμμικό μοντέλο παλινδρόμησης.*

	$\hat{\beta}_1$		$\hat{\beta}_2$		$\hat{\beta}_5$	
Method	*SD*	SD_m (SD_{mad})	*SD*	SD_m (SD_{mad})	*SD*	SD_m (SD_{mad})
SCAD	.167	.171 (.018)	.185	.176 (.022)	.165	.155 (.020)
LASSO	.158	.165 (.022)	.159	.167 (.020	.182	.154 (.019)
Hard	.179	.168 (.018)	.176	.176 (.025)	.157	.154 (.020)
Best subset	.198	.172 (.023)	.185	.175 (.024)	.199	.152 (.023)
Oracle	.163	.199 (.040)	.156	.202 (.043)	.166	.177 (.037)

***Πίνακας 1.3.4.3.4**: Τυπικές αποκλίσεις των εκτιμητών για το εύρωστο γραμμικό μοντέλο παλινδρόμησης.*

Προσομοίωση 3-Λογιστική παλινδρόμηση: Δημιουργήθηκαν 100 σύνολα δεδομένων αποτελούμενα από 200 παρατηρήσεις, βάσει του μοντέλου

$$Y \sim Bernoulli\{p(\underset{\sim}{x}'\underset{\sim}{\beta})\},$$

όπου

$$p(u) = \frac{\exp(u)}{1+\exp(u)},$$

με τις πρώτες 6 συνιστώσες των $\underset{\sim}{\beta}$ και $\underset{\sim}{x}$ να είναι οι ίδιες με αυτές της πρώτης προσομοίωσης. Οι δύο τελευταίες συνιστώσες του $\underset{\sim}{x}$ ήταν i.i.d. από την *Bernoulli* κατανομή με πιθανότητα επιτυχίας 0.5. Επίσης, όλες οι μεταβλητές ήταν κανονικοποιημένες. Τα σφάλματα του μοντέλου υπολογίσθηκαν μέσω 1000 *Monte Carlo* προσομοιώσεων. Τα αποτελέσματα βρίσκονται στους πίνακες 1.3.4.3.5 και 1.3.4.3.6. Η εκτιμήτρια ποινικοποιημένης πιθανοφάνειας με χρήση της *SCAD* είχε καλύτερη απόδοση από αυτήν της *LASSO* και της *Hard*. Επιπλέον, είχε παρόμοια απόδοση συγκριτικά με τον *oracle* εκτιμητή όσον αφορά το *MRME* και την ακρίβεια των εκτιμώμενων τυπικών σφαλμάτων.

		Avg. No. of 0 Coefficients	
Method	*MRME (%)*	*Correct*	*Incorrect*
SCAD ($a = 3.7$)	26.48	4.98	.04
LASSO	53.14	3.76	0
Hard	59.06	4.27	0
Best subset	31.63	4.84	.01
Oracle	25.71	5	0

***Πίνακας 1.3.4.3.5*: Αποτελέσματα προσομοίωσης για τη λογιστική παλινδρόμηση.**

	$\hat{\beta}_1$		$\hat{\beta}_2$		$\hat{\beta}_5$	
Method	*SD*	SD_m (SD_{mad})	*SD*	SD_m (SD_{mad})	*SD*	SD_m (SD_{mad})
SCAD ($a = 3.7$)	.571	.538 (.107)	.383	.372 (.061)	.432	.398 (.065)
LASSO	.310	.379 (.037)	.285	.284 (.019)	.244	.287 (.019)
Hard	.675	.561 (.126)	.428	.400 (.062)	.467	.421 (.079)
Best subset	.624	.547 (.121)	.398	.383 (.067)	.468	.412 (.077)
Oracle	.553	.538 (.103)	.374	.373 (.060)	.432	.398 (.064)

***Πίνακας 1.3.4.3.6*: Τυπικές αποκλίσεις των εκτιμητών για τη λογιστική παλινδρόμηση.**

Παρατηρούμε ότι οι εκτιμώμενες τυπικές αποκλίσεις για τον L_1 εκτιμητή ποινικοποιημένης πιθανοφάνειας (*LASSO*) είναι μικρότερες από αυτές της *SCAD*, αλλά με το συνολικό *MRME* μεγαλύτερο. Αυτό σημαίνει ότι η μεροληψία των εκτιμητών της *LASSO* είναι μεγάλη. Κάτι που ισχύει και για όλες τις προαναφερθείσες προσομοιώσεις.

1.3.5 Συμπεράσματα

Οι μέθοδοι που πρότειναν οι Fan και Li, αποδεδειγμένα έχουν πολύ καλή απόδοση όσον αφορά την επιλογή σημαντικών μεταβλητών. Ο *sandwich* τύπος που κατασκεύασαν για την εκτίμηση των τυπικών σφαλμάτων είναι επίσης αρκετά αποτελεσματικός και ο αλγόριθμος υλοποίησης της όλης μεθόδου υποστηρίζεται από στατιστική θεωρία, με αποτέλεσμα οι εκτιμητές που κατασκευάζονται να έχουν καλές στατιστικές ιδιότητες. Σε σύγκριση με τη μέθοδο επιλογής καλύτερου υποσυνόλου, η οποία είναι αρκετά χρονοβόρα, οι νέες μέθοδοι δίνουν αποτελέσματα αρκετά πιο γρήγορα. Το μεγάλο πλεονέκτημά τους είναι η ταυτόχρονη επιλογή σημαντικών μεταβλητών και η εκτίμηση των συντελεστών, κάτι που γίνεται βελτιστοποιώντας μια ποινικοποιημένη πιθανοφάνεια. Αυτό έχει ως αποτέλεσμα και την ακριβής εκτίμηση των τυπικών σφαλμάτων. Επίσης, απέδειξαν ότι η συνάρτηση ποινής *SCAD*, έχει την καλύτερη απόδοση στην επιλογή σημαντικών μεταβλητών, χωρίς να δημιουργείται μεροληψία, εν αντιθέσει με τη *LASSO* μέθοδο του Tibshirani (1996) όπου χρησιμοποιείται η L_1 συνάρτηση ποινής.

ΚΕΦΑΛΑΙΟ 2

Ποινικοποιημένη παλινδρόμηση με ποινή βασισμένη στη συσχέτιση των μεταβλητών

2.1 Εισαγωγή

Επικεντρωνόμαστε στο σύνηθες γραμμικό μοντέλο παλινδρόμησης

$$y = {}_0 + \mathbf{x}^T\boldsymbol{\beta} + £$$

όπου $\mathbf{x}^T = \left(x_1, ..., x_p\right)$ είναι ένα διάνυσμα μεταβλητών και $£$ είναι ένα διάνυσμα θορύβου με $E\left(£\right) = 0$. Ειδικά για υψηλών διαστάσεων διάνυσμα μεταβλητών $\mathbf{x}$, ο εκτιμητής των συνήθων ελαχίστων τετραγώνων μπορεί να μην είναι μοναδικός. Επιπλέον, δεν είναι η πρώτη επιλογή όταν σκοπός μας είναι η πρόβλεψη. Εναλλακτικοί εκτιμητές όπως της παλινδρόμησης κορυφογραμμής, ridge regression estimator (Hoerl & Kennard, 1970), είναι καλύτεροι και είναι μοναδικοί για μία κατάλληλα επιλεγμένη μειούμενη παράμετρο.

Την τελευταία δεκαετία έχουν προταθεί αρκετοί εναλλακτικοί εκτιμητές που μειώνουν τον αριθμό των παραμέτρων στο τελικό μοντέλο, ειδικά η LASSO (Tibshirani, 1996) που επιβάλλει μια L_1-ποινή στους συντελεστές παλινδρόμησης. Χρησιμοποιώντας μια μη-κυρτή ποινή, κάνει αυτόματα επιλογή μεταβλητών σε αντίθεση με τη παλινδρόμηση κορυφογραμμής που μόνο μειώνει τους εκτιμητές κοντά στο μηδέν. Πιο πρόσφατα, προτάθηκε η μέθοδος elastic net, Enet (Zou & Hastie, 2005), ως μια εναλλακτική διαδικασία που αντιμετωπίζει τις ελλείψεις της LASSO και της παλινδρόμησης κορυφογραμμής, συνδυάζοντας τις L_1 και L_2 ποινές. Ένα κίνητρο των Zou & Hastie ήταν ότι η μέθοδός τους έχει την ιδιότητα να περιλαμβάνει στο τελικό μοντέλο τις ομάδες των μεταβλητών που είναι ισχυρά συσχετισμένες. Όταν οι μεταβλητές είναι ισχυρά συσχετισμένες, η LASSO επιλέγει μόνο μία μεταβλητή από το γκρουπ, ενώ η Enet επιλέγει όλη την ομάδα.

Σε αυτή την εργασία μια εναλλακτική διαδικασία ποινικοποίησης προτείνεται που σκοπεύει στην επιλογή των ομάδων των συσχετισμένων μεταβλητών. Στην απλούστερη έκδοση βασίζεται σε μία

ποινή που κατηγορηματικά χρησιμοποιεί τη συσχέτιση μεταξύ των μεταβλητών σαν βάρη. Στην επεκταμένη έκδοση χρησιμοποιούνται ενισχυτικές τεχνικές για τις ομάδες των μεταβλητών.

2.2 Ποινικοποιημένη παλινδρόμηση συνδεδεμένη με τη συσχέτιση

Έστω τα δεδομένα $\left(y_i, \mathbf{x}_i\right), i = 1, ..., n$, με τα y_i να δηλώνουν την απόκριση και τα $\mathbf{x}_i^T = \left(x_{i1}, ..., x_{ip}\right)$ να δηλώνουν την επεξηγηματική μεταβλητή. Για λόγους απλοποίησης θεωρούμε την απόκριση και τις μεταβλητές κεντραρισμένες. Οι ποινικοποιημένοι εκτιμητές του παραμετρικού διανύσματος $\boldsymbol{\beta}^T = \left(\ _1, ..., \ _p \right)$ μπορούν να αποκομιστούν από την ελαχιστοποίηση των ποινικοποιημένων ελαχίστων τετραγώνων

$$PLS = \sum_{i=1}^{n} \left| y_i - \mathbf{x}_i^T \boldsymbol{\beta} \right|^2 + P(\boldsymbol{\beta}) \quad (1)$$

όπου $P(\boldsymbol{\beta})$ είναι ένας συγκεκριμένος όρος ποινής. Οι κοινές συναρτήσεις ποινής είναι τύπου bridge (Frank & Friedman, 1993, Fu, 1998), δηλαδή

$$P(\boldsymbol{\beta}) = \text{a} \sum_{j=1}^{p} \left| \ _j \right|^{¥}, ¥ > 0,$$

όπου a είναι μία ρυθμιστική παράμετρος. Για $¥ = 2$ λαμβάνουμε τη παλινδρόμηση κορυφογραμμής, για $¥ = 1$ τη LASSO. Οι συναρτήσεις με $¥ < 1$ έχουν ονομαστεί soft περιοριστικές (Donoho & Johnstone, 1995, Klinger, 1998). Η πιο πρόσφατα (το paper γράφτηκε το 2006) προτεινόμενη, Enet, βασίζεται στο συνδυασμό των ποινών που χρησιμοποιούν η LASSO και η παλινδρόμηση κορυφογραμμής, χρησιμοποιώντας έναν όρο ποινής με δύο ρυθμιστικές παραμέτρους a_1, a_2, που δίνεται από τον τύπο

$$P\left(\boldsymbol{\beta}\right) = \text{a}_1 \sum_{j=1}^{p} \left| \ _j \right| + \text{a}_2 \sum_{j=1}^{p} \ _j^2 .$$

Η μέθοδος κληρονομεί ιδιότητες της LASSO πραγματοποιώντας επιλογή μεταβλητών, αλλά σε καταστάσεις όπου η παλινδρόμηση κορυφογραμμής (rigde regression) λειτουργεί καλύτερα ($n > p$ και υψηλή συσχέτιση μεταξύ των μεταβλητών), βασίζεται στον δικό της όρο ποινής. Η Enet τείνει να συμπεριλάβει όλες τις υψηλά συσχετισμένες μεταβλητές, παρά να επιλέξει κάποιες από αυτές.

2.2.1 Εκτιμητής βασισμένος στη συσχέτιση

Η μέθοδος που προτείνεται εδώ αξιοποιεί τη συσχέτιση μεταξύ των μεταβλητών με σαφήνεια στην συνάρτηση ποινής. Οι συντελεστές που αντιστοιχούν σε ζευγάρια μεταβλητών βαρύνονται σύμφωνα με τη συσχέτιση μόνο μεταξύ των ζευγαριών. Η συνάρτηση ποινής δίνεται από τον τύπο

$$P_c(\boldsymbol{\beta}) = \lambda \sum_{i=1}^{p-1} \sum_{j>i} \left\{ \frac{\left(\beta_i - \beta_j \right)^2}{1-\rho_{ij}} + \frac{\left(\beta_i + \beta_j \right)^2}{1+\rho_{ij}} \right\} \quad (2)$$
$$= 2\lambda \sum_{i=1}^{p-1} \sum_{j>i} \frac{\beta_i^2 - 2\rho_{ij}\beta_i\beta_j + \beta_j^2}{1-\rho_{ij}^2}$$

όπου ρ_{ij} δηλώνει τη συσχέτιση μεταξύ των i και j μεταβλητών. Έχει σχεδιαστεί με τρόπο ώστε για ισχυρή θετική συσχέτιση ($\rho_{ij} \to 1$) ο πρώτος όρος να γίνεται επικρατέστερος, έχοντας ως αποτέλεσμα οι εκτιμητές των β_i, β_j να είναι παρόμοιοι ($\hat{\beta}_i \approx \hat{\beta}_j$). Για ισχυρή αρνητική συσχέτιση ($\rho_{ij} \to -1$) ο δεύτερος όρος γίνεται επικρατέστερος και το $\hat{\beta}_i$ πλησιάζει το $-\hat{\beta}_j$. Το αποτέλεσμα είναι η ομαδοποίηση, δηλαδή οι ισχυρές συσχετίσεις δείχνουν συγκρίσιμες τιμές εκτιμητών ($\left|\hat{\beta}_i\right| \approx \left|\hat{\beta}_j\right|$), με το πρόσημο να καθορίζεται από τη θετική ή αρνητική συσχέτιση. Όταν οι μεταβλητές είναι ασυσχέτιστες λαμβάνουμε (ανάλογα με μια σταθερά) τη ποινή της παλινδρόμησης κορυφογραμμής $P_c(\boldsymbol{\beta}) \propto \lambda \sum \beta_\S^2$. Συνεπώς, για αδύναμη συσχέτιση μεταξύ των δεδομένων το αποτέλεσμα είναι αρκετά κοντά στον εκτιμητή της παλινδρόμησης κορυφογραμμής. Επομένως, όπως στην Enet η παλινδρόμηση κορυφογραμμής είναι μία οριακή περίπτωση.

Ένα καλό χαρακτηριστικό της ποινής (2) είναι ότι μπορεί να δοθεί σε απλή τετραγωνική μορφή, το οποίο επιτρέπει να δώσει τον προκύπτων εκτιμητή σε κλειστή μορφή. Λαμβάνουμε

$$P_c(\boldsymbol{\beta}) = \lambda \boldsymbol{\beta}^T \mathbf{W} \boldsymbol{\beta}$$

όπου $\mathbf{W}$ είναι ένας πίνακας που καθορίζεται από τις συσχετίσεις $\rho_{ij}, i, j = 1, \ldots, p$ (για λεπτομέρειες πάνω στο $\mathbf{W}$ βλέπε επόμενη ενότητα). Για $\rho_{ij}^2 \neq 1, \lambda > 0$, μια άμεση λύση του προβλήματος ποινικοποιημένων ελαχίστων τετραγώνων (1) λαμβάνεται από τον εκτιμητή βασισμένο στη συσχέτιση

$$\hat{\boldsymbol{\beta}}_c = \left(\mathbf{X}^T \mathbf{X} + \lambda \mathbf{W} \right)^{-1} \mathbf{X}^T \mathbf{y} \text{ , (3)}$$

όπου $\mathbf{X}^T = (\mathbf{x}_1, ..., \mathbf{x}_n)$ είναι ο πίνακας σχεδιασμού και $\mathbf{y}$ είναι το διάνυσμα των αποκρίσεων, $\mathbf{y}^T = (y_1, ..., y_n)$.

2.2.2 Κατασκευή της ποινής

Το αποτέλεσμα της ομαδοποίησης εξαρτάται ισχυρά από την κυρτότητα του όρου ποινής. Η ποινή που βασίζεται στη συσχέτιση μπορεί να ιδωθεί ως ένας συνδυασμός δύο ποινών, $P_c(\boldsymbol{\beta}) = P_{c,1}(\boldsymbol{\beta}) + P_{c,2}(\boldsymbol{\beta})$ όπου

$$P_{c,1}(\boldsymbol{\beta}) = \alpha \sum_i \sum_{j>i} \frac{(\beta_i - \beta_j)^2}{1 - \rho_{ij}},$$

$$P_{c,2}(\boldsymbol{\beta}) = \alpha \sum_i \sum_{j>i} \frac{(\beta_i + \beta_j)^2}{1 + \rho_{ij}}.$$

Ο πρώτος όρος γίνεται σημαντικός για θετικά συσχετισμένες μεταβλητές όταν ο άλλος όρος είναι σημαντικός για αρνητικά συσχετισμένες μεταβλητές. Ούτε ο $P_{c,1}(.)$ ούτε ο $P_{c,2}(.)$ είναι αυστηρώς κυρτός. Άλλα (για $\alpha > 0$ και $\rho_{ij}^2 \neq 1$ αν $i \neq j$) ο συνδυασμός $P_c(\boldsymbol{\beta})$ είναι αυστηρώς κυρτός. Μία ωραία συνέπεια αυτού είναι ότι ο εκτιμητής $\hat{\boldsymbol{\beta}}_c$ υπάρχει και είναι μοναδικός.

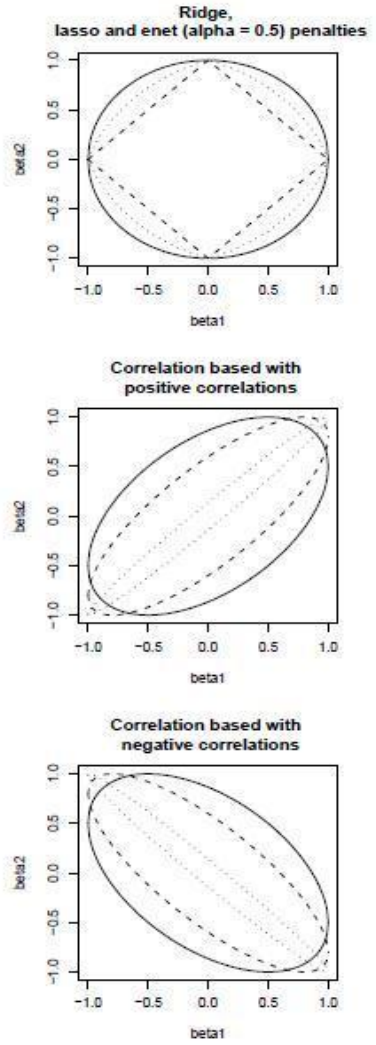
Πρόταση 2.2.2.1

Έστω ότι $\alpha > 0$ και $\rho_{ij}^2 \neq 1$ για $i \neq j$. Τότε ισχύουν τα ακόλουθα

1) Ο όρος $P_c(\boldsymbol{\beta})$ είναι αυστηρώς κυρτός.
2) Ο εκτιμητής $\hat{\boldsymbol{\beta}}_c$ υπάρχει και είναι μοναδικός.
3) Ο όρος $P_c(\boldsymbol{\beta})$ μπορεί να δοθεί σε τετραγωνική μορφή $P_c(\boldsymbol{\beta}) = \alpha \boldsymbol{\beta}^T \mathbf{W} \boldsymbol{\beta}$ όπου ο πίνακας $\mathbf{W} = (w_{ij})$ καθορίζεται από

$$w_{ij} = \begin{cases} 2\sum_{s \neq i} \frac{1}{1 - \rho_{is}^2}, & i = j, \\ -2\frac{\rho_{ij}}{1 - \rho_{ij}^2}, & i \neq j. \end{cases} \quad (4)$$

(για την απόδειξη βλέπε παράρτημα). Έτσι για $\alpha > 0$ ο εκτιμητής που βασίζεται στη συσχέτιση μοιράζεται την ιδιότητα της ύπαρξης και μοναδικότητας με τον εκτιμητή της παλινδρόμησης κορυφογραμμής. Σε αντίθεση, ο LASSO εκτιμητής δεν έχει κατ' ανάγκη μία μοναδική λύση.



Γράφημα 1: *Πάνω πίνακας: Ποινικοποιημένα ελάχιστα τετράγωνα με ρ=0(συμπαγής γραμμή), LASSO ποινή(παύλες) και Enet(τελείες). Μεσαίος πίνακας: Τρεις τιμές θετικής συσχέτισης: ρ=0.5(συμπαγής), ρ=0.8(παύλες), και ρ=0.99(τελείες). Κάτω πίνακας: Τρεις τιμές αρνητικής συσχέτισης: ρ=-0.5(συμπαγής), ρ=-0.8(παύλες), και ρ=-0.99(τελείες).*

Το γράφημα 1 δείχνει τα δισδιάστατα περιγράμματα για επιλεγμένες τιμές του °. Η περιοχή που περιορίζεται η ποινή της παλινδρόμησης κορυφογραμμής είναι ο δίσκος $^{2}_{1} + ^{2}_{2} \leq c$, για τη LASSO ο ρόμβος $\left| _{1} \right| + \left| _{2} \right| \leq c$. Όσο ο ρόμβος έχει ευδιάκριτες γωνίες, αν μία λύση πετυχαίνεται σε μία γωνία, τότε μία παράμετρος $_{j}$ γίνεται ίση με μηδέν. Φαίνεται ότι τα περιγράμματα της παλινδρόμησης κορυφογραμμής και της LASSO είναι αρκετά συμμετρικά, ο $x_1 = 0$ είναι άξονας συμμετρίας τόσο καλά όσο ο $x_2 = 0$. Αντίθετα, η περιοχή που περιορίζεται ο βασισμένος στη συσχέτιση εκτιμητής είναι ένα ελλειψοειδές που γίνεται στενότερο όσο αυξάνεται η συσχέτιση. Η

ανάλυση του πίνακα $P_c(\boldsymbol{\beta})$ αποφέρει τα ιδιοδιανύσματα $(1,1)$ και $(1,-1)$ με αντίστοιχες ιδιοτιμές $\alpha/(1-\circ)$ και $\alpha/(1+\circ)$. Έτσι για $\circ > 0$ η πρώτη ιδιοτιμή υπερισχύει ενώ για $\circ < 0$ είναι η δεύτερη ιδιοτιμή που καθορίζει τον προσανατολισμό του ελλειψοειδούς. Καθώς υπολογίζουμε τον ποινικοποιημένο εκτιμητή ελαχίστων τετραγώνων το αποτέλεσμα είναι ότι για $\circ > 0$ οι εκτιμητές που προτιμώνται είναι αυτοί με όμοιες τιμές, ενώ για $\circ < 0$ προτιμάται η ομοιότητα του τύπου $\hat{\beta}_i \approx -\hat{\beta}_j$. Αυτό μπορεί να φανεί και τα περιγράμματα, όπου για $\circ > 0$ η αύξηση του $P_c(\boldsymbol{\beta})$ είναι πιο αργή όταν κινείται στη διεύθυνση του πρώτου ιδιοδιανύσματος $(1,1)$ από ότι στην ορθογώνια διεύθυνση $(1,-1)$. Για $\circ < 0$ η ιδιοτιμή που αντιστοιχεί στο $(1,-1)$ είναι μεγαλύτερη και επομένως προτιμούνται οι τιμές των παραμέτρων εκείνων που το β_1 είναι κοντά στο $-\beta_2$. Έτσι η χρήση της ποινής P_c συνεπάγεται μείωση, η οποία μείωση καθορίζεται από το α, αλλά η μείωση αυτή διαφέρει από αυτή της παλινδρόμησης κορυφογραμμής, η οποία πετυχαίνεται στην ειδική περίπτωση $\circ_{ij} = 0$.

2.2.3 Το αποτέλεσμα της ομαδοποίησης: Η ακραία περίπτωση

Μια μέθοδος παλινδρόμησης εμφανίζει το αποτέλεσμα της ομαδοποίησης αν οι συντελεστές παλινδρόμησης μιας ομάδας υψηλά συσχετισμένων μεταβλητών τείνουν να είναι ίσοι (κατ' απόλυτη τιμή). Για τη γενική μέθοδο ποινικοποίησης (1) έχει δειχθεί ότι για ταυτόσημα διανύσματα μεταβλητών $x_i = x_j$, αυτό συνεπάγεται $\hat{\beta}_i = \hat{\beta}_j$, αν η $P(\boldsymbol{\beta})$ είναι αυστηρώς κυρτή (Λήμμα 2, Zou & Hastie, 2005). Παρ' όλα αυτά, για τον εκτιμητή που βασίζεται στη συσχέτιση

$$\hat{\boldsymbol{\beta}}_c = \arg\min_{\boldsymbol{\beta}} |\mathbf{y} - \mathbf{X}\boldsymbol{\beta}|^2 + P_c(\boldsymbol{\beta}) \quad (5)$$

η άμεση λύση $\hat{\boldsymbol{\beta}}_c = \left(\mathbf{X}^T\mathbf{X} + P_c(\boldsymbol{\beta})\right)^{-1}\mathbf{X}^T\mathbf{y}$ είναι διαθέσιμη μόνο για μη τέλεια συσχετισμένες μεταβλητές. Αν $x_i = x_j$, τότε η συνάρτηση ποινής δεν είναι πλέον αυστηρώς κυρτή και το Λήμμα 2 των Zou & Hastie, (2005) δεν εφαρμόζεται. Παρ' όλα αυτά, αν και για $\circ_{ij}^2 \to 1$ η συνάρτηση $P_c(\boldsymbol{\beta})$ χειροτερεύει, η εκτίμηση μπορεί να οριστεί ως το όριο. Με $\hat{\boldsymbol{\beta}}_c\left(\alpha, \{\circ_{ij}\}\right)$ να δηλώνεται η λύση της (5), ορίζουμε για $\circ_{ij}^2 = 1$ τον εκτιμητή που βασίζεται στη συσχέτιση ως

$$\hat{\boldsymbol{\beta}}_c\left(\alpha, \{\circ_{ij}\}\right) = \lim_{\circ_{ij}^2 \to 1} \boldsymbol{\beta}_c\left(\alpha, \{\sigma_{ij}\}\right)$$

όπου το όριο το παίρνουμε για $\sigma_{ij} \to 1$ αν $x_i = x_j$ και $\sigma_{ij} \to -1$ αν $x_i = -x_j$. Για πρακτικούς καθαρά σκοπούς βρήκαμε οτι η τιμή $\sigma_{ij} = 0.98$ δουλεύει καλά για την οριακή εκτίμηση. Για καλύτερη

επεξήγηση, εξετάζουμε την ειδική περίπτωση $p=2$ πιο στενά. Έστω $P_c(\boldsymbol{\beta}) = {}^{\text{a}}\boldsymbol{\beta}^T\mathbf{W}\boldsymbol{\beta} = {}^{\text{a}}\boldsymbol{\beta}^T\mathbf{D}_2^T\mathbf{D}_2\boldsymbol{\beta}$ όπου

$$\mathbf{D}_2 = \begin{pmatrix} 1/\sqrt{1-\text{°}} & -1/\sqrt{1-\text{°}} \\ 1/\sqrt{1+\text{°}} & 1/\sqrt{1+\text{°}} \end{pmatrix},$$
$$\mathbf{W} = \frac{2}{1-\text{°}^2}\begin{pmatrix} 1 & -\text{°} \\ -\text{°} & 1 \end{pmatrix}.$$

Για την οριακή περίπτωση $\text{°}^2 \to 1$ η αντιστροφή μπορεί να υπολογιστεί άμεσα. Έχουμε,

$$\lim_{\text{°}\to 1}\left(\mathbf{X}^T\mathbf{X} + {}^{\text{a}}\mathbf{W}\right)^{-1} = \frac{1}{2\left(2+{}^{\text{a}}\right)}\begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix},$$
$$\lim_{\text{°}\to -1}\left(\mathbf{X}^T\mathbf{X} + {}^{\text{a}}\mathbf{W}\right)^{-1} = \frac{1}{2^{\text{a}}}\begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}.$$

Είναι άμεσα εμφανές ότι στο όριο λαμβάνουμε $\hat{}_1 = \hat{}_2$ για $\text{°} = 1$ και $\hat{}_1 = -\hat{}_2$ για $\text{°} = -1$. Επιπλέον, στην ειδική περίπτωση $p=2$ οι ιδιοτιμές των $\mathbf{X}^T\mathbf{X}$ και $\mathbf{W}$ είναι οι ίδιες και μπορεί να δειχθεί οτι, δοσμένης της ύπαρξης του εκτιμητή ελαχίστων τετραγώνων $\hat{\boldsymbol{\beta}}_{OLS}$, ο εκτιμητής που βασίζεται στη συσχέτιση είναι μία μειούμενη εκδοχή του, ήτοι $\hat{\boldsymbol{\beta}}_c = ¥\hat{\boldsymbol{\beta}}_{OLS}$, όπου $¥ = \left(1-\text{°}^2\right)/\left(1-\text{°}^2+2^{\text{a}}\right)$. Αν $\text{°} \neq 0$, αυτό είναι διαφορετικό από την παλινδρόμηση κορυφογραμμής όπου η συρρίκνωση είναι αναφορική στην ορθοκανονική βάση που παράγεται από τις στήλες του $\mathbf{X}$ (Hastie et al., 2001, Κεφάλαιο 3).

2.3 Προσομοιώσεις σε μεσαίων διαστάσεων προβλήματα

Ακολούθως, πρώτα ερευνούμε την απόδοση διαφόρων μεθόδων για μεσαίου μεγέθους αριθμό μεταβλητών. Οι προσομοιώσεις είναι παρόμοιες με αυτές που χρησιμοποιήθηκαν στο paper της lasso (Tibshirani, 1996) και στο paper της Enet (Zou & Hastie, 2005). Το βασικό μοντέλο παλινδρόμησης δίνεται από τη σχέση

$$\mathbf{y} = \mathbf{x}^{\text{T}}\boldsymbol{\beta} + \pm£, \quad £ \sim N(0,1).$$

Κάθε σετ δεδομένων αποτελείται από ένα σύνολο εκπαίδευσης (training set), πάνω στο οποίο προσαρμόστηκαν τα μοντέλα, ένα σύνολο επικύρωσης (validation set), το οποίο χρησιμοποιήθηκε για την επιλογή των ρυθμιστικών παραμέτρων, και ένα σύνολο ελέγχου (test set) για την αξιολόγηση της απόδοσης. Ο συμβολισμός $\cdot||\cdot$ χρησιμοποιείται για να περιγράψει τον αριθμό των παρατηρήσεων στα σύνολα εκπαίδευσης, επικύρωσης και ελέγχου, αντίστοιχα. Στις προσομοιώσεις κεντράρουμε τις

μεταβλητές με βάση το σύνολο εκπαίδευσης. Έστω $\bar{\mathbf{x}}_{train} = \left(\bar{x}_{1,train}, \ldots, \bar{x}_{p,train}\right)$ να ορίζει το διάνυσμα των μέσων από το σύνολο εκπαίδευσης, n_{test} ο αριθμός των παρατηρήσεων στο σύνολο ελέγχου και $\bar{y}_{train}$ ο μέσος των αποκρίσεων στο σύνολο εκπαίδευσης.

Χρησιμοποιούμε δύο μέτρα απόδοσης, το τεστ σφάλματος (mean squared error) $MSE_y = \frac{1}{n_{test}} \mathbf{r}_{sim}^T \mathbf{r}_{sim}$ με $r_{i,sim} = \mathbf{x}_i^T \boldsymbol{\beta} - \left(\bar{y}_{train} + \left(\mathbf{x}_i - \bar{\mathbf{x}}_{train}\right)^T \hat{\boldsymbol{\beta}}\right)$ στο σύνολο ελέγχου και το μέσο τετραγωνικό σφάλμα για την εκτίμηση του $\boldsymbol{\beta}$, $MSE = \left|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right|^2$. Όσο το πρώτο μέτρο αξιολογεί την προβλεπτική ικανότητα του μοντέλου, το δεύτερο στοχεύει στην ακρίβεια του εκτιμητή και επομένως στον προσδιορισμό της επίδρασης των μεταβλητών.

Ακολουθούν οι περιπτώσεις-σενάρια που ερευνήθηκαν, προσομοιώνοντας 50 σετ δεδομένων.

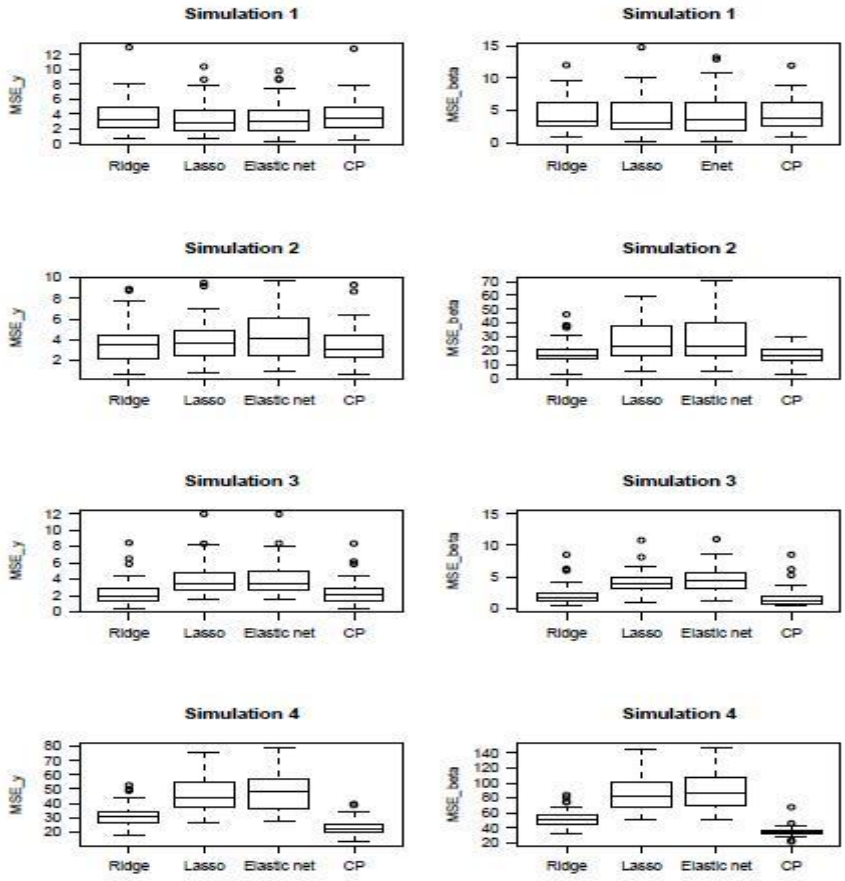
1) Στην πρώτη περίπτωση με $p = 8$, το $\boldsymbol{\beta}$ ορίζεται να είναι $\boldsymbol{\beta}^T = (3, 1.5, 0, 0, 2, 0, 0, 0)$ και $\pm = 3$. Η ανά ζεύγη συσχέτιση ορίζεται να είναι $^{\circ}\left(x_i, x_j\right) = 0.5^{|i-j|}$. Το μέγεθος του δείγματος είναι $20|20|200$.
2) Με $p = 9$, το $\boldsymbol{\beta}$ ορίζεται να είναι $\boldsymbol{\beta}^T = (1, 2, 3, 4, 0, 4, 3, 2, 1)$ και $\pm = 3$, $^{\circ}\left(x_i, x_j\right) = 1 - 0.25|i - j|$, ίδιο μέγεθος δείγματος όπως στο (1).
3) Ίδιο με το (1) εκτός από το οτι $_1 = \ _2 = \ldots = \ _8 = 0,85$.
4) Με $p = 40$, το διάνυσμα των συντελεστών είναι $\boldsymbol{\beta}^T = (\underbrace{0, \ldots, 0}_{10}, \underbrace{2, \ldots, 2}_{10}, \underbrace{0, \ldots, 0}_{10}, \underbrace{2, \ldots, 2}_{10})$, $\pm = 15$, $^{\circ}\left(x_i, x_j\right) = 0.5$, για όλα τα i και j. Το μέγεθος του δείγματος είναι $100|100|400$.

Method	Simulation 1		Simulation 2		Simulation 3		Simulation 4	
	median MSE_y	median MSE_β	median MSE_y	median MSE_β	median MSE_y	median MSE_β	median MSE_y	median MSE_β
Ridge Regr.	3.28	3.35	3.57	17.12	**1.91**	1.69	30.25	51.67
LASSO	**2.92**	**3.13**	3.73	28.83	3.35	3.99	43.69	83.26
Elastic Net	2.96	3.65	4.15	23.91	3.46	4.44	47.95	87.64
CP	3.40	3.68	**3.06**	**16.95**	2.07	**1.20**	**21.95**	**34.47**

Πίνακας 1: *Διάμεσος των δύο τεστ για τα προσομοιωμένα παραδείγματα, μετά από 50 επαναλήψεις.*

Τα αποτελέσματα των προσομοιώσεων δίνονται στους Πίνακα 1 και Γράφημα 2. Στον Πίνακα 1 η καλύτερη απόδοση σημειώνεται με έντονη γραφή. Το πρώτο παράδειγμα περιλαμβάνει μόνο θετικά

συσχετισμένες μεταβλητές ενώ στο δεύτερο υπάρχουν και θετικά και αρνητικά συσχετισμένες. Τα παραδείγματα 3 και 4 περιλαμβάνουν ομάδες μεταβλητών: στο 3 υπάρχει μόνο ένα γκρουπ, ενώ στο 4 υπάρχουν δύο γκρουπ σχετικών μεταβλητών. τα παραδείγματα 1, 3 και 4 αντιστοιχούν στα 1, 2 και 3 των Zou & Hastie (2005). Φαίνεται οτι στην πρώτη περίπτωση των θετικά συσχετισμένων μεταβλητών η Enet λειτουργεί καλύτερα, αλλά οι μέθοδοι δεν διαφέρουν σημαντικά στην απόδοσή τους. Στην περίπτωση θετικά και αρνητικά συσχετισμένων μεταβλητών, υπερέχει η εκτίμηση που βασίζεται στη συσχέτιση. Κυρίως στο τελευταίο παράδειγμα με την ύπαρξη ομάδων, ο εκτιμητής CP υπερέχει καθαρά από τις υπόλοιπες μεθόδους και στα δύο μέτρα απόδοσης.



Γράφημα 2: *Θηκογραφήματα των δύο τεστ για τις προσομοιώσεις 1-4.*

2.4 Ενίσχυση κατά ομάδες

Ο εκτιμητής που βασίζεται στη συσχέτιση (2) πετυχαίνει συρρίκνωση των συντελεστών, αλλά όχι επιλογή μεταβλητών. Ως εκ τούτου, ειδικότερα σε υψηλών διαστάσεων προβλήματα έχει κάποια μειονεκτήματα. Μία μέθοδος που λειτουργεί πολύ καλά σε μεγάλες διαστάσεις είναι η κατά

συνιστώσα ενίσχυση (componentwise boosting). Προερχόμενη από την κοινότητα μηχανικής μάθησης έχει δειχθεί ότι έχει καλές ιδιότητες στην παλινδρόμηση (Bühlmann & Yu, 2003, Bühlmann, 2006).

Για να ξεπεράσουμε τα μειονεκτήματα του CP προτείνουμε ένα νέο τρόπο κατά συνιστώσα ενίσχυσης. Με σκοπό να εξασφαλίσουμε το αποτέλεσμα της ομαδοποίησης του CP σε συνδυασμό με την επιλογή μεταβλητών θεωρούμε μια ενισχυτική διαδικασία που σε κάθε βήμα ανανεώνει τους συντελεστές περισσότερων από μίας μεταβλητής. Αυτή η διαδικασία διαφέρει από τις κοινές διαδικασίες κατά συνιστώσα ενίσχυσης, όπου σε κάθε βήμα επιλέγεται μία μεταβλητή και προσαρμόζεται ο αντίστοιχος συντελεστής. Για να διαχωρίσουμε την κατά συνιστώσα ενίσχυση από τη διαδικασία που προτείνεται εδώ θα αναφερόμαστε πλέον σε αυτήν ως ενίσχυση κατά ομάδες (blockwise boosting).

Έστω $S \subset \{1,...,p\}$ το σύνολο των δεικτών των μεταβλητών που θεωρούνται σε ένα δοσμένο βήμα. Η βασική ιδέα είναι να υπολογίσουμε σε ένα βήμα της επαναληπτικής διαδικασίας τις παραμέτρους που ελαχιστοποιούν το κριτήριο των ποινικοποιημένων ελαχίστων τετραγώνων

$$\left|\mathbf{r} - \mathbf{X}_S \mathbf{b}\right|^2 + P_{c,S}\,, \qquad (6)$$

όπου το $\mathbf{r}$ ορίζεται το διάνυσμα των υπολοίπων (από το προηγούμενο βήμα), $\mathbf{X}_S$ είναι ο μειωμένος πίνακας σχεδιασμού που περιλαμβάνει μόνο τις μεταβλητές $j \in S$ και $P_{c,S}$ η ποινή βασισμένη στη συσχέτιση του υποσυνόλου S. Με την ελαχιστοποίηση της (6) παίρνουμε μία ταυτόχρονη αναπροσαρμογή για τα στοιχεία του S. Όπως συνήθως στις ενισχυτικές διαδικασίες, σε κάθε βήμα χρησιμοποιήθηκε ένα ασθενές κριτήριο. Αυτό σημαίνει οτι μόνο μικρή αλλαγή στην εκτίμηση της παραμέτρου μπορεί να συμβεί σε κάθε βήμα. Ως εκ τούτου η παράμετρος ^{a} στην (6) επιλέγεται πολύ μεγάλη, στην περίπτωσή μας $^{a} \geq 1000$. Έχει δειχθεί (Bühlmann & Yu, 2003) οτι μεγάλες τιμές της ^{a} αποφέρουν καλύτερες αποδόσεις. Ο μόνος περιορισμός είναι το υπολογιστικό κόστος, αφού πολύ μεγάλες τιμές της ^{a} απαιτούν πολλά επαναληπτικά βήματα.

Όπως στην κατά συνιστώσα ενίσχυση, η επιλογή μεταβλητών επιτυγχάνεται επιλέγοντας σε κάθε βήμα ένα κατάλληλο υποσύνολο S. Θεωρώντας όλα τα πιθανά υποσύνολα, αυτό συνεπάγεται μεγάλο υπολογιστικό κόστος, ακόμα και μικρό αριθμό μεταβλητών. Ως εκ τούτου τα υποψήφια σύνολα μειώνονται κατατάσσοντας τις μεταβλητές (όπως στην κατά συνιστώσα ενίσχυση) και έπειτα θεωρώντας ως υποψήφια σύνολα μόνο τα υποσύνολα του S που δημιουργούνται από την επιτυχή εισαγωγή μιας μεταβλητής από τη δοσμένη κατάταξη. Έτσι σε κάθε βήμα φτιάχνεται μία κατάταξη των μεταβλητών.

Για τα υποσύνολα που περιλαμβάνουν μόνο μία μεταβλητή, η συνάρτηση P_c δεν μπορεί να χρησιμοποιηθεί άμεσα. Σε αυτές τις περιπτώσεις χρησιμοποιούμε την ποινή της παλινδρόμησης κορυφογραμμής $P_{c,\{j\}} = \text{a}\ _j^2$. Η καταλληλότητα των υποσυνόλων εκτιμάται από ένα κριτήριο βασισμένο στην πληροφορία, το AIC (Akaike Information Criterion), το οποίο επίσης χρησιμοποιείται ως κριτήριο τερματισμού. Πρώτα παρουσιάζουμε τον αλγόριθμο, και μετά παραθέτουμε τον προσδιορισμό του AIC που χρησιμοποιήσαμε.

Αλγόριθμος

Βήμα 1: (Αρχικοποίηση)

Θέτουμε $\hat{\boldsymbol{\beta}}^{(0)} = 0, \hat{\boldsymbol{\mu}}^{(0)} = 0.$

Βήμα 2: (Επανάληψη)

Για $m = 1, 2, \ldots$

a) Κατατάσσουμε κατάλληλα τις μεταβλητές σύμφωνα με το πόσο βελτιώνουν την προσαρμογή του μοντέλου

Υπολογίζουμε τα υπόλοιπα $\mathbf{r}^{(m)} = \mathbf{y} - \hat{\boldsymbol{\mu}}^{(m-1)}$ και για $j \in \{1, \ldots, p\}$ προσαρμόζουμε το μοντέλο $\mathbf{r}^{(m)} = \mathbf{X}_{\{j\}} b_j + \boldsymbol{\varepsilon}$ ελαχιστοποιώντας την ποσότητα $\left|\mathbf{r}^{(m)} - \mathbf{X}_{\{j\}} b_j\right|^2 + P_{c,\{j\}}$, παίρνοντας τα $\hat{b}_{j1}, \ldots, \hat{b}_{jp}$ έτσι ώστε $AIC(\hat{b}_{j1}) \le \ldots \le AIC(\hat{b}_{jp})$.

b) Βρίσκουμε τον κατάλληλο αριθμό μεταβλητών για ανανέωση

Για $r = 1, \ldots, p$

Με $S_r = \{j_1, \ldots, j_r\}$ προσαρμόζουμε το μοντέλο $\mathbf{r}^{(m)} = \mathbf{X}_{S_r} \mathbf{b}_{S_r} + \boldsymbol{\varepsilon}$ ελαχιστοποιώντας την ποσότητα $\left|\mathbf{r}^{(m)} - \mathbf{X}_{S_r} \mathbf{b}_{S_r}\right|^2 + P_{c,S_r}$, παίρνοντας τους εκτιμητές $\hat{\mathbf{b}}_{S_r}$ και τα $AIC(\hat{\mathbf{b}}_{S_r})$.

c) Επιλογή

Επιλέγουμε το υποσύνολο των μεταβλητών που έχει την καλύτερη προσαρμογή

$$S^{(m)} = \arg\min_{S_r} AIC(\hat{\mathbf{b}}_{S_r}).$$

d) Αναπροσαρμογή

Το διάνυσμα των παραμέτρων ανανεώνεται ως εξής:

$$\hat{\beta}_j^{(m)} = \begin{cases} \hat{\beta}_j^{(m-1)} + \hat{b}_j, & \text{αν } j \in S^{(m)}, \\ \hat{\beta}_j^{(m-1)}, & \text{αλλιώς,} \end{cases}$$

παίρνοντας το διάνυσμα $\hat{\boldsymbol{\beta}}^{(m)} = \left(\hat{\beta}_1^{(m)}, \ldots, \hat{\beta}_p^{(m)} \right)^T$ και $\hat{\boldsymbol{\mu}}^{(m)} = \hat{\boldsymbol{\mu}}^{(m-1)} + \mathbf{X}_{S^{(m)}} \hat{\mathbf{b}}_{S^{(m)}}$.

Το κριτήριο τερματισμού που προτείνουμε είναι μία έκδοση του AIC, $AIC = -2\left(l\left(\hat{\boldsymbol{\mu}}^{(m)}\right) - tr\left(\mathbf{H}_m\right) \right)$ όπου $l\left(\hat{\boldsymbol{\mu}}^{(m)}\right)$ ορίζεται η λογαριθμοποιημένη πιθανοφάνεια μετά την mth προσαρμογή και $tr\left(\mathbf{H}_m\right)$ το ίχνος του αντίστοιχου πίνακα προβολής. Κάποιες πράξεις δείχνουν ότι δίνεται από τη σχέση

$$\hat{\boldsymbol{\mu}}^{(m)} = \mathbf{H}_m \mathbf{y}$$

όπου

$$\begin{aligned} \mathbf{H}_m &= \sum_{j=1}^{m} \tilde{\mathbf{H}}_m \prod_{l=1}^{j-1} \left(I - \tilde{\mathbf{H}}_{j-l} \right) \\ &= \tilde{\mathbf{H}}_1 + \tilde{\mathbf{H}}_2 \left(I - \tilde{\mathbf{H}}_1 \right) + \ldots \end{aligned}$$

με $\tilde{\mathbf{H}}_j = \mathbf{X}_{S^{(j)}} \left(\mathbf{X}_{S^{(j)}}^T \mathbf{X}_{S^{(j)}} + \alpha \mathbf{W}_{S^{(j)}} \right)^{-1} \mathbf{X}_{S^{(j)}}^T$, όπου $\mathbf{W}_{S^{(j)}}$ ορίζεται ο πίνακας ποινής από τη σχέση (3) για τα υποσύνολα $S^{(j)}$.

Χρησιμοποιούμε το διορθωμένο κριτήριο AIC (Hurvich et al., 1998) με ένα πρόσθετο διορθωτικό παράγοντα

$$AIC_c = \log\left(\hat{\sigma}_m^2 \right) + \frac{1 + 1.8 \cdot tr\left(\mathbf{H}_m\right)/n}{1 - \left(1.8 \cdot tr\left(\mathbf{H}_m\right) + 2\right)/n},$$

όπου

$$\hat{\sigma}_m^2 = \frac{1}{n} \left(\mathbf{y} - \hat{\boldsymbol{\mu}}^{(m)} \right)^T \left(\mathbf{y} - \hat{\boldsymbol{\mu}}^{(m)} \right).$$

Ο εκτιμητής $\hat{\boldsymbol{\beta}}_B$ που λαμβάνουμε από τη διαδικασία BlockBoost κληρονομεί την ισχυρή ιδιότητα της ομαδοποίησης των μεταβλητών από τον εκτιμητή που βασίζεται στη συσχέτιση. Αν οι μεταβλητές είναι ισχυρά συσχετισμένες οι αντίστοιχες ανανεώσεις τους στον αλγόριθμο έχουν κατ' απόλυτη τιμή παρόμοιες τιμές.

Μέσα στον αλγόριθμο ο εκτιμητής που βασίζεται στη συσχέτιση χρησιμοποιείται για υποσύνολα διαφόρων μεγεθών. Η ρυθμιστική παράμετρος α που χρησιμοποιείται πρέπει να προσαρμόζεται ανάλογα με τον αριθμό των μεταβλητών. Για υποσύνολα που περιέχουν μία μεταβλητή η ρυθμιστική παράμετρος είναι α. Για μεγαλύτερα υποσύνολα χρησιμοποιούμε την ποινή

$$P_{c,S}(\boldsymbol{\beta}) = \alpha_{|S|} \sum_{\substack{i<j \\ (i,j)\in S}} \left\{ \frac{(\beta_i - \beta_j)^2}{1-\rho_{ij}} + \frac{(\beta_i + \beta_j)^2}{1+\rho_{ij}} \right\} = 2\alpha_{|S|} \sum_{\substack{i<j \\ (i,j)\in S}} \frac{\beta_i^2 - 2\rho_{ij}\beta_i\beta_j + \beta_j^2}{1-\rho_{ij}^2} \quad (7)$$

όπου $\alpha_{|S|}$ είναι η ρυθμιστική παράμετρος που εξαρτάται μόνο από την πληθικότητα του S, $|S|$. Με σκοπό να έχουμε μόνο μία ρυθμιστική παράμετρο η $\alpha_{|S|}$ επιλέγεται σαν μία συνάρτηση της α. Αν θεωρήσουμε την περίπτωση των ασυσχέτιστων μεταβλητών η ποινή για τις μεταβλητές που δίνεται από την (2) γίνεται $P_c(\boldsymbol{\beta}) = 2\alpha(p-1)\sum_{i=1}^{p}\beta_i^2$ όπου είναι η ποινή της παλινδρόμησης κορυφογραμμής με παράμετρο $2\alpha(p-1)$. Ως εκ τούτου η $\alpha_{|S|}$ στην (7) επιλέγεται ως $\alpha_{|S|} = \alpha(|S|-1)$.

Προτού ερευνήσουμε την απόδοση σε υψηλών διαστάσεων προβλήματα, αποδεικνύουμε το αποτέλεσμα της ομαδοποίησης σε μία μικρή προσομοίωση και εξετάζουμε την επιλογή μεταβλητών σε πραγματικά δεδομένα.

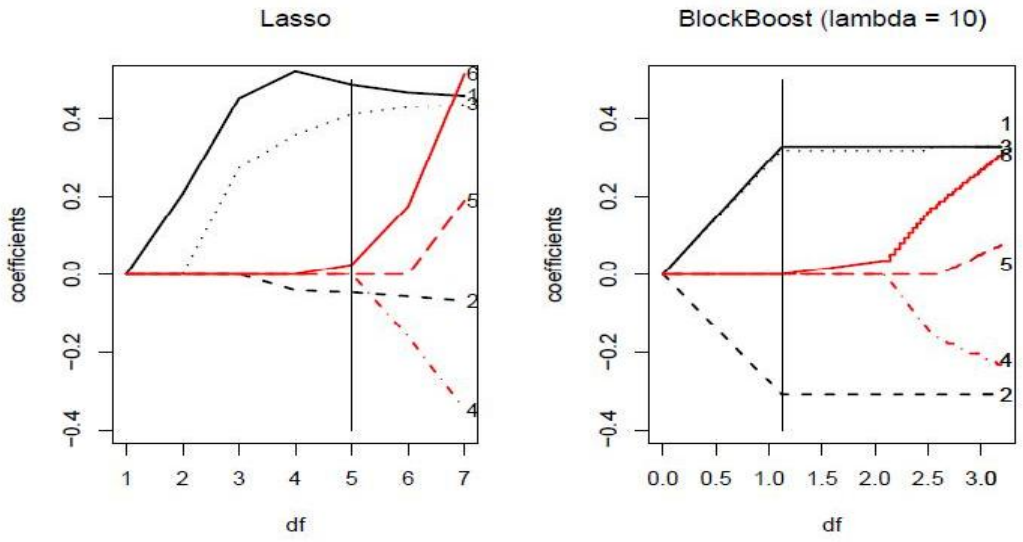
2.4.1 Το αποτέλεσμα της ομαδοποίησης

Για την επεξήγηση του αποτελέσματος της ομαδοποίησης παίρνουμε ένα εξιδανικευμένο παράδειγμα από τους Zou & Hastie (2005). Με Z_1 και Z_2 να είναι δύο ανεξάρτητες μεταβλητές από την $U(0,20)$ η απόκριση παράγεται από την $N(Z_1 + 0.1Z_2, 1)$. Υποθέτουμε οτι οι παρατηρήσεις είναι

$$\mathbf{x}_1 = Z_1 + \varepsilon_1,\ \mathbf{x}_2 = -Z_1 + \varepsilon_2,\ \mathbf{x}_3 = Z_1 + \varepsilon_3,$$
$$\mathbf{x}_4 = Z_2 + \varepsilon_4,\ \mathbf{x}_5 = -Z_2 + \varepsilon_5,\ \mathbf{x}_6 = Z_2 + \varepsilon_6,$$

όπου ε_i είναι ανεξάρτητες και ισόνομες μεταβλητές από την $N(0,1/16)$. Οι μεταβλητές x_1, x_2 και x_3 μπορεί να θεωρηθεί οτι φτιάχνουν μία ομάδα, ενώ οι x_4, x_5 και x_6 μία δεύτερη. Το Γράφημα 3 δείχνει

την κατασκευή των συντελεστών για τη Lasso και τη BlockBoost για μέγεθος δείγματος $n=100$. Φαίνεται οτι η BlockBoost ομαδοποιεί τις μεταβλητές x_1, x_2 και x_3 και οι αντίστοιχες εκτιμήτριές τους είναι κατ' απόλυτη τιμή όμοιες. Το οτι αποτελούν ομάδα προσδιορίζεται σαφώς. Η Lasso δείχνει μερικώς διαφορετική κατασκευή των συντελεστών επιλέγοντας ως ισχυρά δεμένες τις x_1 και x_3, και με πιο αδύναμη συσχέτιση τη x_2. Ενώ οι συντελεστές της BlockBoost αντανακλούν την ισχυρή συσχέτιση των x_1, x_2 και x_3, οι της Lasso είναι μάλλον ακανόνιστοι. Η μέθοδος Enet συμπεριφέρεται σχεδόν παρόμοια με τη BlockBoost (σύγκρινε με Zou & Hastie, 2005).



Γράφημα 3: *Κατασκευή συντελεστών για Lasso (αριστερά) και BlockBoost (δεξιά)*

2.4.2 Εφαρμογή σε πραγματικά δεδομένα (σωματικό λίπος)

Το σετ δεδομένων για το σωματικό λίπος χρησιμοποιήθηκε από τους Penrose et al. (1985). Η μελέτη σκοπεύει στην εκτίμηση του ποσοστού του σωματικού λίπους από διάφορες μετρήσεις σε περιφέρειες του σώματος σε 252 άντρες. Οι δεκατρείς μεταβλητές είναι: 1) ηλικία, 2) βάρος (λίβρες), 3) ύψος (ίντσες), 4) περιφέρεια λαιμού, 5) περιφέρεια στήθους, 6) κοιλιακή περιφέρεια, 7) περιφέρεια ισχίου, 8) περιφέρεια μηρού, 9) περιφέρεια γονάτου, 10) περιφέρεια αστραγάλου, 11) περιφέρεια δικέφαλου μυ (επεκταμένου), 12) περιφέρεια πήχη, και 13) περιφέρεια καρπού. Όλες οι περιφέρειες μετρήθηκαν σε cm. Το ποσοστό του σωματικού λίπους υπολογίστηκε από την εξίσωση του Siri (1956), χρησιμοποιώντας την πυκνότητα του σώματος καθορισμένη σε υποβρύχια ζύγιση.

Με σκοπό να ερευνήσουμε τις επιδόσεις των εναλλακτικών μεθόδων 'σπάσαμε' το σετ δεδομένων 20 φορές με τυχαίο τρόπο σε ένα σύνολο εκπαίδευσης 151 παρατηρήσεων και ένα σύνολο

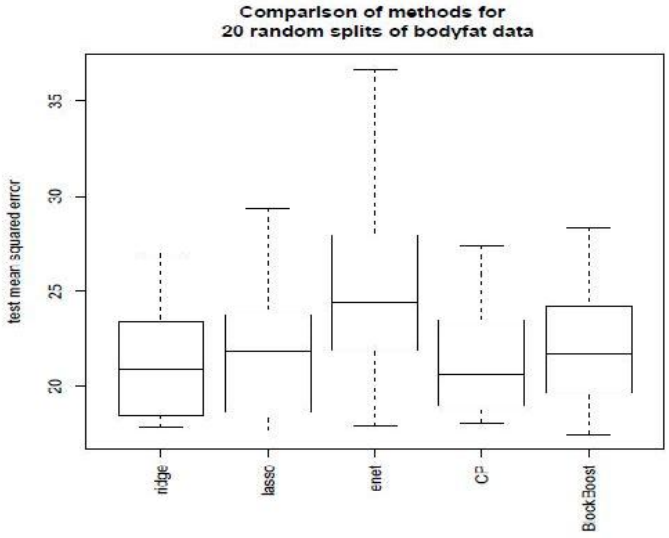
ελέγχου 101 παρατηρήσεων. Οι ρυθμιστικές παράμετροι επιλέχθηκαν με τη διαδικασία της δεκαπλής διασταυρωμένης επικύρωσης.

Η απόδοση ως προς τη διάμεσο των μέσων τετραγωνικών σφαλμάτων δίνεται στον Πίνακα 2, τα αντίστοιχα θηκογραφήματα στο Γράφημα 4. Φαίνεται οτι ο εκτιμητής που βασίζεται στη συσχέτιση έχει την καλύτερη απόδοση ως προς το μέσο τετραγωνικό σφάλμα, οι BlockBoost και Enet επιλέγουν τον ίδιο αριθμό μεταβλητών.

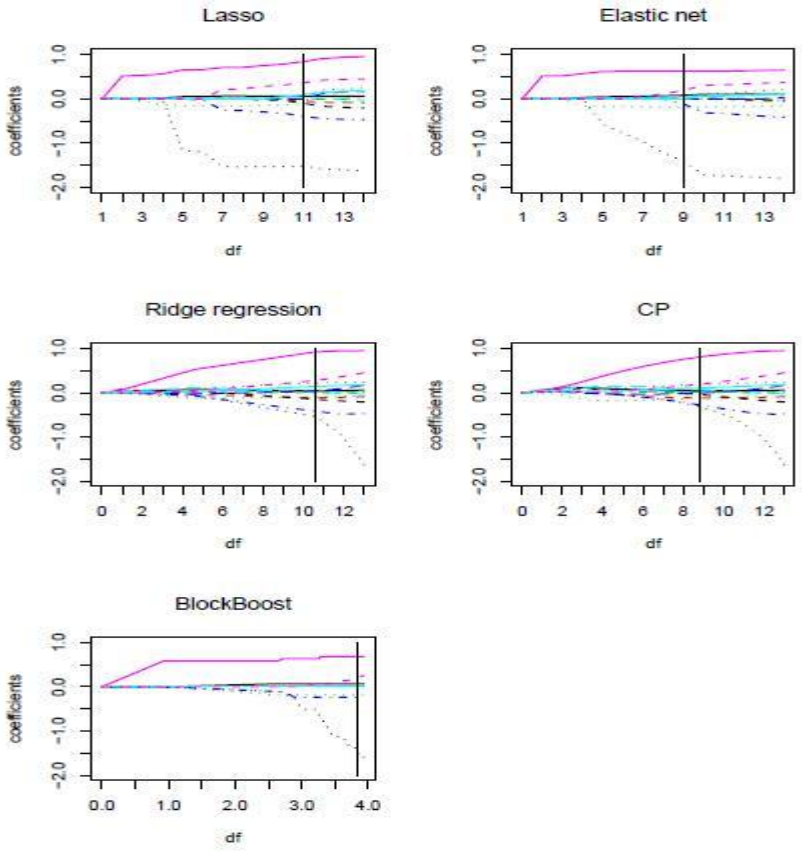
Το Γράφημα 5 απεικονίζει την κατασκευή των συντελεστών για Lasso, Enet, παλινδρόμηση κορυφογραμμής, CP και BlockBoost βασισμένοι σε όλο το σετ δεδομένων. Φαίνεται οτι η παλινδρόμηση κορυφογραμμής και η εκτίμηση με βάση τη συσχέτιση λειτουργούν παρόμοια. Υπάρχει επίσης μια ομοιότητα μεταξύ της Lasso και της Enet. Η BlockBoost επιλέγει 5 μεταβλητές ενώ οι Lasso και Enet επιλέγουν 11 και 9, αντίστοιχα. Η μείωση στις σχετικές μεταβλητές είναι περίπου η ίδια για τις δύο διαδικασίες ενώ η BlockBoost περιλαμβάνει λιγότερες μεταβλητές στο τελικό μοντέλο. Οι εκτιμήτριες που δίνονται στον Πίνακα 3 δείχνουν οτι μεγάλες διαφορές βρίσκονται μόνο στις μεταβλητές 4, 12 και 13.

Method	median MSE_y	median no. of selected variables
Ridge regression	20.84	13
Lasso	21.80	9.5
Elastic net	24.38	6
CP	20.67	13
BlockBoost	21.70	6

Πίνακας 2: Σωματικό λίπος- διάμεσοι MSE μετά από 20 τυχαία 'σπασίματα' του σετ δεδομένων.



Γράφημα 4: *Θηκογραφήματα των διαμέσων των MSE.*



Γράφημα 5: *Κατασκευή συντελεστών για τις 5 μεθόδους.*

Variables	Ridge	Lasso	Elastic net	CP	BlockBoost
Tuning parameters:	$\lambda = 148.41$	$s = 0.79$	$\lambda = 0.05$ $s = 0.77$	$\lambda = 8.17$	$\lambda = 2000$ $m = 56$
1	0.07	0.06	0.09	0.08	0.09
2	-0.03	-0.05	0	-0.03	0
3	-0.16	-0.11	-0.19	-0.18	-0.15
4	-0.43	-0.40	-0.24	-0.28	-0.11
5	0.05	0	0.06	0.1	0
6	0.77	0.86	0.63	0.65	0.69
7	-0.16	-0.11	0	-0.03	0
8	0.19	0.12	0.09	0.13	0
9	0.10	0	0	0.08	0
10	0.02	0.02	0	-0.002	0
11	-0.04	0.10	0.03	-0.05	0
12	0.01	0.37	0.26	-0.04	0
13	-0.39	-1.53	-1.60	-0.242	-1.29

Πίνακας 3: *Ρυθμιστικές παράμετροι και εκτιμήτριες για το συνολικό σετ δεδομένων.*

2.5 Απόδοση σε συνθήκες υψηλών διαστάσεων

2.5.1 Προβλεπτική ικανότητα και εκτίμηση των επιδράσεων

Στα ακόλουθα χρησιμοποιούμε τον ίδιο συμβολισμό όπως στις προσομοιώσεις της Ενότητας 2.3. Ωστόσο, τώρα επικεντρωνόμαστε σε υψηλών διαστάσεων προβλήματα με πολλές μεταβλητές. Ακολουθούν τα τρία σενάρια που χρησιμοποιούμε.

1) Το διάνυσμα των παραμέτρων στο πρώτο σενάριο είναι $\boldsymbol{\beta}^T = (\underbrace{3,...,3}_{5}, \underbrace{0,...,0}_{5}, \underbrace{3,...,3}_{5}, \underbrace{0,...,0}_{5}, \underbrace{3,...,3}_{5}, \underbrace{0,...,0}_{25})$. Η συσχέτιση $° \left(x_i, x_j\right) = °_{ij}$ δίνεται από τη σχέση

$$°_{ij} = \begin{cases} 1 - 0.1 \cdot |i-j|, & i, j \in \{k, k+1, ..., k+4\}, k \in \{1, 6, 11, 16, 21\} \\ £_{ij}, & \text{αλλιώς,} \end{cases}$$

όπου $£_{ij}$ είναι iid από την $N\left(0, 0.1^2\right)$. Το μέγεθος του δείγματος είναι $20|20|40$.

2) Παρόμοιες συνθήκες με το σενάριο 1, με ασθενέστερη συσχέτιση οριζόμενη ως

$$°_{ij} = 1 - 0.05 \cdot |i-j|, \quad i, j \in \{k, k+1, ..., k+4\}, k \in \{1, 6, 11, 16, 21\}.$$

3) Το διάνυσμα είναι

$$\boldsymbol{\beta}^T = (5, 4, 3, 2, 1, \underbrace{0,...,0}_{5}, 5, 4, 3, 2, 1, \underbrace{0,...,0}_{5}, -5, -4, -3, -2, -1, \underbrace{0,...,0}_{25})$$

και η συσχέτιση ορίζεται

$$°_{ij} = \begin{cases} 1 - 0.075 \cdot |i-j|, & i, j \in \{k, k+1, ..., k+4\}, k \in \{1, 6, 11, 16, 21\} \\ £_{ij}, & \text{αλλιώς,} \end{cases}$$

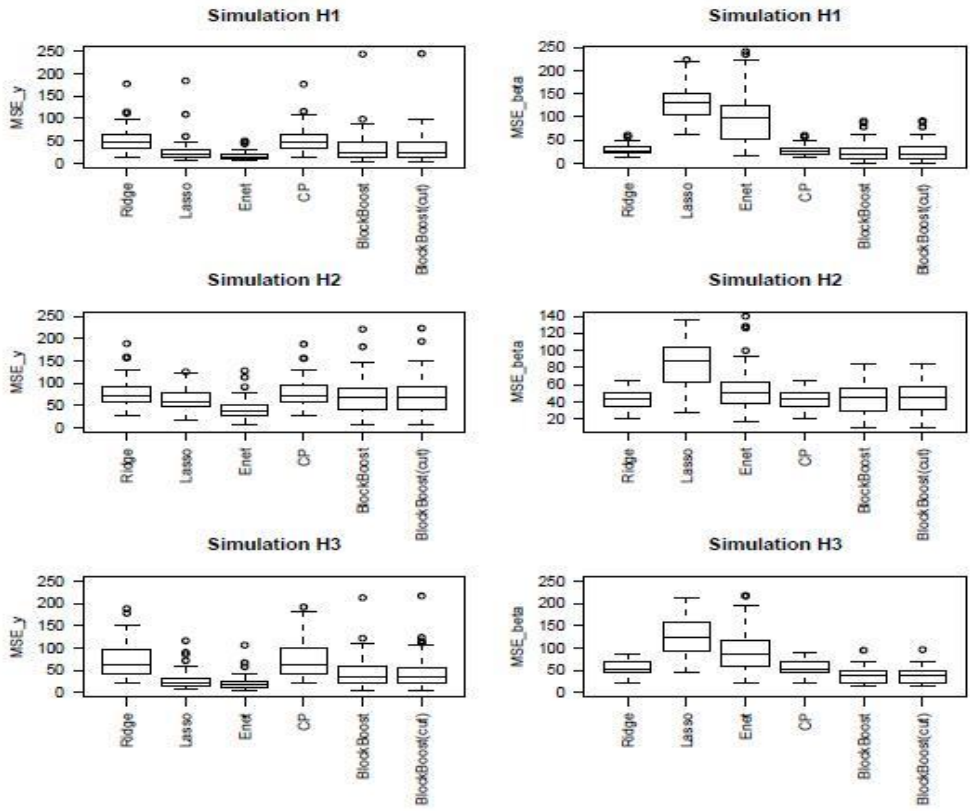
όπου $£_{ij}$ είναι iid από την $N\left(0, 0.1^2\right)$. Το μέγεθος του δείγματος είναι $20|20|40$.

Στις προσομοιώσεις ερευνούμε μία πρόσθετη παραλλαγή της BlockBoost, την BlockBoost(cut). Είναι σχεδιασμένη να διαγράφει τις μεταβλητές που επιλέχθηκαν μία ή δύο φορές στην επαναληπτική διαδικασία. Πιο συγκεκριμένα, η μεταβλητή i διαγράφεται, αν $\left| \ _i \middle/ \sum_j \right| \ _j \Big| < 0.01$. Θεωρούμε ως μέτρα σύγκρισης της απόδοσης των μεθόδων τα MSE_y και MSE .

Τα αποτελέσματα των προσομοιώσεων δίνονται στους Πίνακα 4 και Γράφημα 6. Στα τρία παραδείγματα η Enet έχει την καλύτερη πρόβλεψη, ακολουθούμενη από τη Lasso. Ωστόσο, αν εξετάσουμε την ακρίβεια της εκτίμησης της παραμέτρου, η απόδοση της BlockBoost είναι σαφώς ανώτερη από αυτές των Enet και Lasso. Η BlockBoost φαίνεται να κυριαρχεί όσον αφορά το MSE . Ένας λόγος είναι οτι η BlockBoost κάνει κάτι καλύτερο στον προσδιορισμό σημαντικών μεταβλητών. Αυτό το αποτέλεσμα εξετάζεται πιο στενά στην επόμενη υποενότητα.

Method	Simulation H1		Simulation H2		Simulation H3	
	median MSE_y	median MSE_β	median MSE_y	median MSE_β	median MSE_y	median MSE_β
Ridge Regression	48.11	27.27	72.77	44.33	63.35	54.06
LASSO	19.62	131.63	59.86	88.06	20.90	125.51
Elastic Net	**14.94**	97.28	**38.55**	51.31	**17.34**	87.84
CP	47.63	26.48	72.99	**44.32**	64.08	53.96
BlockBoost	23.01	**19.36**	67.10	44.99	36.01	38.11
BlockBoost (cut)	23.53	19.67	69.54	45.15	35.02	**38.00**

Πίνακας 4: *Διάμεσοι των MSE_y και MSE για τα παραδείγματα 1-3, βασισμένα σε 50 επαναλήψεις.*



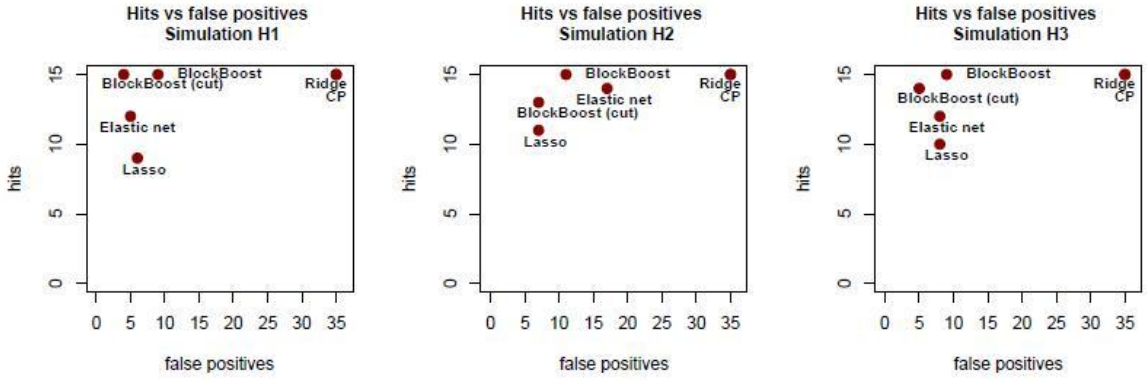
Γράφημα 6: *Θηκογραφήματα των MSE_y και MSE για τα παραδείγματα 1-3.*

2.5.2 Προσδιορισμός σημαντικών μεταβλητών

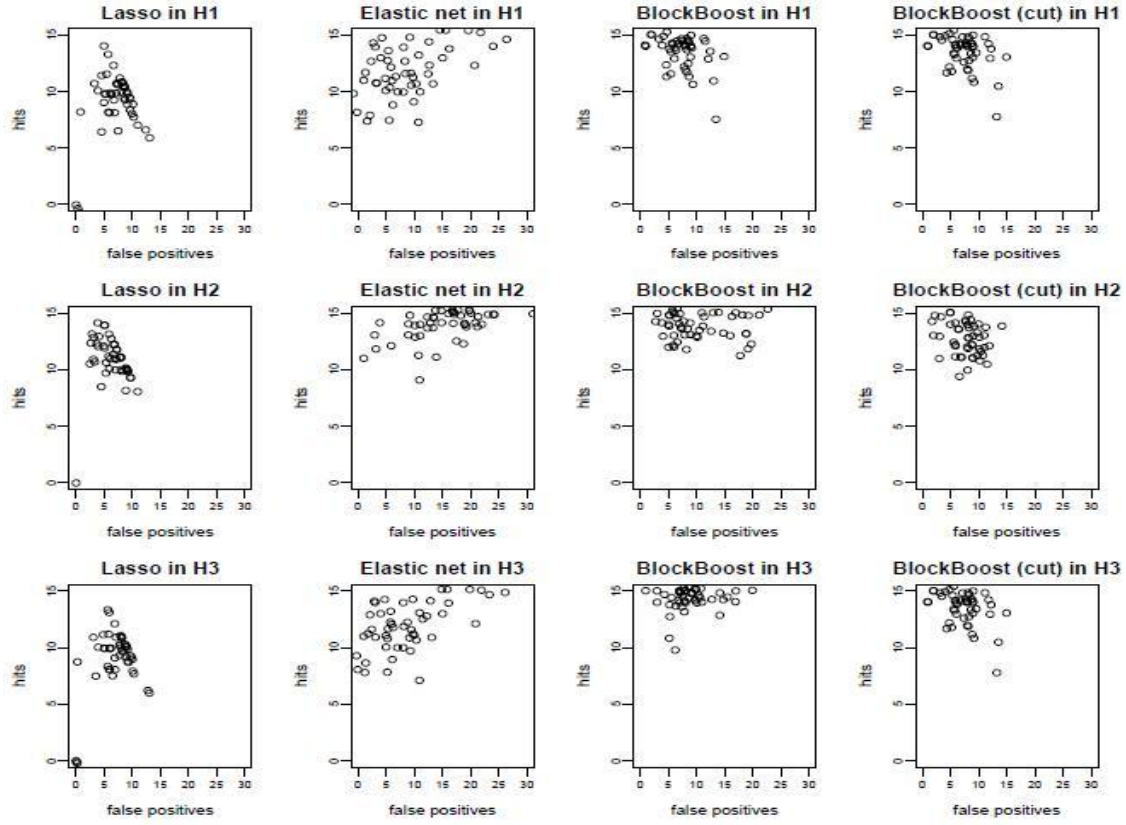
Ενώ η προβλεπτική ικανότητα είναι ένα σημαντικό κριτήριο σύγκρισης των μεθόδων, αυτό που ενδιαφέρει περισσότερο τους πρακτικούς ερευνητές είναι οι μεταβλητές που περιλαμβάνονται στο τελικό μοντέλο. Το τελικό μοντέλο πρέπει να είναι όσο το δυνατό πιο οικονομικό, αλλά πρέπει να περιλαμβάνει όλες τις σημαντικές μεταβλητές. Τα κριτήρια με τα οποία κρίνεται η απόδοση των διαδικασιών είναι οι 'επιτυχίες' (ο αριθμός των σωστά προσδιορισμένων σημαντικών μεταβλητών) και οι 'αποτυχίες' (ο αριθμός των μη σημαντικών μεταβλητών που κρίθηκαν σημαντικές). Ο Πίνακας 5 και το Γράφημα 7 δείχνουν τις μέσες τιμές των 'επιτυχιών' και 'αποτυχιών' για τα σενάρια που προηγήθηκαν. Το Γράφημα 7 κατασκευάστηκε με τρόπο παρόμοιο με τα ROC κύματα, μόνο αντί για κύματα μπήκαν σημεία. Η περίπτωση της ιδανικής απόδοσης αντιστοιχεί στην πάνω αριστερά γωνία. Αποκλίσεις στα δεξιά αντιστοιχούν σε αύξηση των 'αποτυχιών', κακή απόδοση στον προσδιορισμό σημαντικών μεταβλητών αντιστοιχεί σε μικρές τιμές στην τετμημένη. Λόγω της κατασκευής τους, οι εκτιμητές CP και της παλινδρόμησης κορυφογραμμής βρίσκονται στην πάνω δεξιά γωνία που σημαίνει οτι όλες οι σημαντικές μεταβλητές περιλαμβάνονται αλλά και οι μη σημαντικές. Απ' την άλλη η BlockBoost (όπως και η BlockBoost(cut)) λειτουργούν πολύ καλύτερα στον προσδιορισμό των σημαντικών μεταβλητών. Η Lasso σαφώς χάνει κάποιες σημαντικές μεταβλητές, αλλά και η Enet έχει τάση να χάνει κάποιες. Όσον αφορά τις 'αποτυχίες' η BlockBoost είναι συγκρίσιμη με τις Enet και Lasso ενώ η BlockBoost(cut) έχει την καλύτερη απόδοση. Η ίδια τάση φαίνεται και στο Γράφημα 8 όπου οι αποδόσεις των Enet, Lasso και BlockBoost παρουσιάζονται για κάθε ξεχωριστή προσομοίωση.

Method	Results for the following examples:					
	Example H1		Example H2		Example H3	
	hits	false positives	hits	false positives	hits	false positives
Ridge regression	15	35	15	35	15	35
Lasso	9	6	11	7	10	8
Elastic net	12	6	14	16	12	8
CP	15	35	15	35	15	35
BlockBoost	15	9	15	11	15	9
BlockBoost (cut)	15	4	13	7	14	5

Πίνακας 5: *Μέσες τιμές των 'επιτυχιών' και 'αποτυχιών' για τα σενάρια 1-3.*



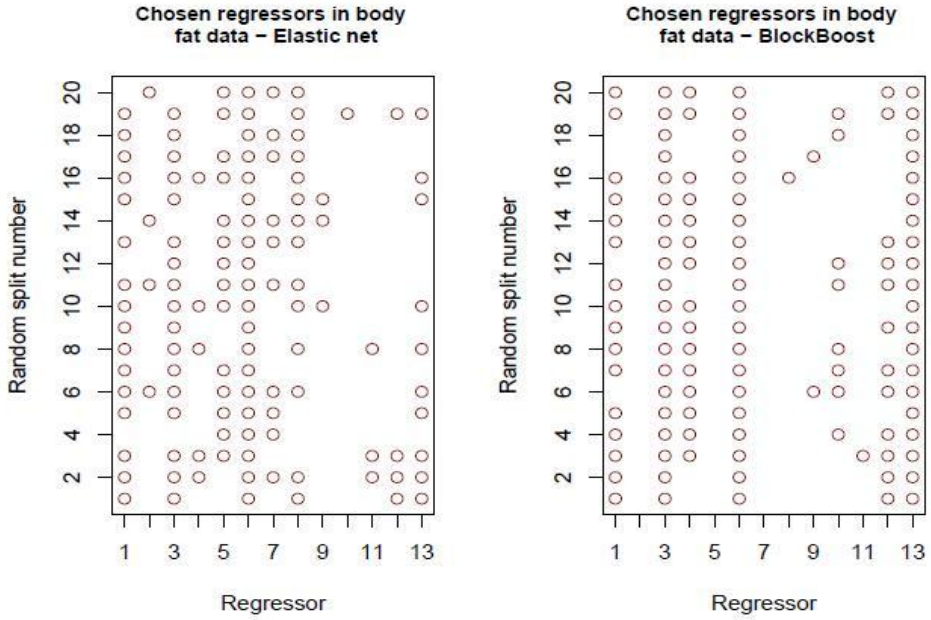
Γράφημα 7: *Μέσες 'επιτυχίες' ενάντια στις μέσες 'αποτυχίες' για τα σενάρια 1-3.*



Γράφημα 8: *'Επιτυχίες' ενάντια στις 'αποτυχίες' για κάθε ξεχωριστή προσομοίωση των σεναρίων 1-3.*

Για την απόδοση σε πραγματικά δεδομένα θεωρούμε πάλι το σετ με το σωματικό λίπος. Το Γράφημα 9 δείχνει τις μεταβλητές που επιλέγουν οι BlockBoost και Enet στις 20 διασπάσεις του σετ δεδομένων. Φαίνεται ότι η BlockBoost έχει λιγότερη μεταβλητότητα στην επιλογή των σημαντικών μεταβλητών. Για παράδειγμα η BlockBoost ποτέ δεν επιλέγει την μεταβλητή 2 ενόσω επιλέγεται σε 4 περιπτώσεις από την Enet. Η μεταβλητή 13 επιλέγεται πάντα από την BlockBoost, μα μόνο στο 50 %

των περιπτώσεων από την Enet. Έστω $h_i, i = 1,...,13$ να είναι ο αριθμός των διασπάσεων που επιλέγουν την i μεταβλητή. Θεωρώντας τα $h_1,...,h_{13}$ ως μετρήσεις μπορούμε να συγκρίνουμε τις τυπικές αποκλίσεις τους. Παρατηρούμε οτι για την BlockBoost είναι 8.55 και για την Enet 6.37, το οποίο δείχνει ότι η BlockBoost είναι πιο ευσταθής υπό την έννοια οτι τείνει να επιλέγει τις ίδιες μεταβλητές μεταξύ των διασπάσεων.



Γράφημα 9: *Σύγκριση των επιλεγμένων μεταβλητών μεταξύ της Enet και της BlockBoost για τις 20 τυχαίες διασπάσεις του σετ δεδομένων του σωματικού λίπους.*

2.6 Συμπεράσματα

Στο Κεφάλαιο αυτό έχουν προταθεί δύο αλγόριθμοι για την προσαρμογή γραμμικών μοντέλων, οι οποίοι όπως και η Enet επικεντρώνονται στο αποτέλεσμα της ομαδοποίησης. Αποδείχθηκε οτι, παρόλο που η Enet έχει πλεονεκτήματα στην προβλεπτική ισχύ, οι αλγόριθμοι που βασίζονται στη συσχέτιση δείχνουν να έχουν θεαματική απόδοση αν η επιτυχία μετριέται από το σωστό προσδιορισμό των σημαντικών μεταβλητών. Δεδομένου οτι στις εφαρμογές, ειδικά στα υψηλών διαστάσεων προβλήματα, ο προσδιορισμός των σημαντικών μεταβλητών είναι αποφασιστικής σημασίας, η μέθοδος μπορεί να θεωρηθεί ως ένας ισχυρός ανταγωνιστής στο πεδίο αυτό. Η μέθοδος μπορεί να επεκταθεί σε γενικευμένα γραμμικά μοντέλα χρησιμοποιώντας μία ποινικοποιημένη προσέγγιση πιθανοφάνειας. Για την ποινή που βασίζεται στη συσχέτιση η επέκταση γίνεται άμεσα. Οι ενισχυτικές μέθοδοι μπορούν να επιτευχθούν τροποποιώντας τη LogitBoost (Friedman et al., 1999).

ΚΕΦΑΛΑΙΟ 3

Ποινικοποιημένη ενισχυμένη παλινδρόμηση με ποινή βασισμένη στη συσχέτιση σε γενικευμένα γραμμικά μοντέλα

3.1 Εισαγωγή

Τα γραμμικά μοντέλα έχουν μία μακρά παράδοση στη στατιστική. Όταν ο αριθμός των μεταβλητών είναι μεγάλος, η εκτίμηση των άγνωστων παραμέτρων συχνά αντιμετωπίζει προβλήματα. Τότε το ενδιαφέρον επικεντρώνεται στην από τα δεδομένα καθοδηγούμενη επιλογή των σημαντικών μεταβλητών. Ο σύγχρονος εξοπλισμός παρακολούθησης ο οποίος χρησιμοποιείται ευρέως σε πολλές διαδικασίες συλλογής δεδομένων κάνει δυνατό το να συλλέγουμε δεδομένα με μεγάλο αριθμό μεταβλητών, ακόμα και με αισθητά περισσότερες επεξηγηματικές μεταβλητές από οτι παρατηρήσεις. Ένα παράδειγμα είναι η ανάλυση των δεδομένων των γονιδιακών εκφράσεων. Εδώ τα τυπικά καθήκοντα είναι η επιλογή μεταβλητών και η ταξινόμηση των δειγμάτων σε δύο ή περισσότερες διαφορετικές κατηγορίες. Οι δυαδικές αποκρίσεις μπορούν να αντιμετωπιστούν στο πλαίσιο των γενικευμένων γραμμικών μοντέλων (Nelder & Wedderburn, 1972) και επίσης εξετάζονται από τον Toutenburg (1992).

Υπάρχουν διάφορες προσεγγίσεις για την επίτευξη επιλογής υποσυνόλου στα γενικευμένα γραμμικά μοντέλα. Μειούμενες μέθοδοι με L_1 ποινές, όπως ο Lasso εκτιμητής, είναι μία κλάση μεθόδων. Ο Lasso εκτιμητής παρουσιάστηκε από τον Tibshirani (1996) για τα γραμμικά μοντέλα και επεκτάθηκε στα γενικευμένα γραμμικά μοντέλα από τους Park & Hastie (2007). Μία διαφορετική προσέγγιση είναι η κατά στοιχείο ενίσχυση (componentwise boosting), (Bühlmann & Yu, 2003). Η κατά στοιχείο ενίσχυση χρησιμοποιεί έναν αδύναμο αλγόριθμο μάθησης για να βελτιώσει τον εκτιμητή. Πετυχαίνουμε επιλογή υποσυνόλου αν σε κάθε βήμα, ο αλγόριθμος είναι περιορισμένος να χρησιμοποιεί ένα υποσύνολο μεταβλητών.

Μία πλευρά στην επιλογή υποσυνόλου, που επισημαίνεται από τους Zou & Hastie (2005), είναι η μεταχείριση υψηλά συσχετισμένων μεταβλητών. Αντί να επιλέγουμε μία αντιπρόσωπο από μία ομάδα υψηλά συσχετισμένων μεταβλητών, μπορούμε να 'ενθαρρύνουμε' αυτές τις μεταβλητές να είναι μέσα ή έξω από το μοντέλο όλες μαζί. Οι Zou & Hastie αναφέρονται σε αυτό ως αποτέλεσμα της ομαδοποίησης.

Σε αυτό το κεφάλαιο παρουσιάζουμε μία νέα μέθοδο ποινικοποίησης και μία ενισχυτική έκδοσή της, η οποία επικεντρώνεται ρητά στην επιλογή των ομάδων. Για να πετύχουμε το στόχο αυτό θεωρούμε μία ποινή η οποία χρησιμοποιεί τη συσχέτιση μεταξύ των μεταβλητών ως βάρη για την ποινικοποίηση. Βλέπε προηγούμενο κεφάλαιο για την παρόμοια προσέγγιση στα γραμμικά μοντέλα. Αυτή η νέα μέθοδος και κάποιες ιδιότητές της περιγράφονται στην ενότητα 2. Μία ενισχυτική έκδοση αυτής, που παρουσιάζεται στην ενότητα 3, επιτρέπει την επιλογή μεταβλητών. Στην ενότητα 4 χρησιμοποιούμε προσομοιωμένα και πραγματικά σετ δεδομένων για να συγκρίνουμε τις νέες μεθόδους με αυτές που ήδη υπάρχουν.

3.2 Ποινικοποιημένη εκτιμήτρια μέγιστης πιθανοφάνειας

Θεωρούμε ένα σύνολο από n ανεξάρτητες μονοδιάστατες παρατηρήσεις $y_1,\ldots,y_n$, με κατανομές πυκνότητας από την απλή εκθετική οικογένεια

$$f\left(y|\bar{},\ddot{}\right)=\exp\left\{\frac{y\bar{}-b(\bar{})}{\ddot{}}+c\left(y,\ddot{}\right)\right\}, \quad (1)$$

όπου $\bar{}$ είναι η φυσική βαθμωτή παράμετρος της οικογένειας, $\ddot{}>0$ είναι η παράμετρος ενόχλησης ή διασποράς, $b(\cdot)$ και $c(\cdot)$ είναι μετρήσιμες συναρτήσεις. Επίσης, για κάθε παρατήρηση, καταγράφεται ένα σύνολο p επεξηγηματικών μεταβλητών $\mathbf{x}_i=\left(x_{i1},\ldots,x_{ip}\right)^T$. Αυτές σχηματίζουν μία γραμμική παράμετρο πρόβλεψης $\dagger_i=\ _0+\mathbf{x}_i^T\boldsymbol{\beta}^*$, όπου η $\ _0$ είναι μία σταθερά και $\boldsymbol{\beta}^*=\left(\ _1,\ldots,\ _p\right)^T$ είναι ένα p διάστατο παραμετρικό διάνυσμα. Θεωρούμε οτι η αναμενόμενη τιμή της y_i δίνεται από τη σχέση $«_i=h\left(\dagger_i\right)$, όπου $h(\cdot)$ είναι μία διαφορίσιμη μονότονη συνάρτηση και $«_i$ είναι η αναμενόμενη τιμή της y_i.

Θεωρώντας οτι η παράμετρος ενόχλησης $\ddot{}$ είναι γνωστή, ενδιαφερόμαστε να βρούμε το άγνωστο παραμετρικό διάνυσμα $\boldsymbol{\beta}=\left(\ _0,\boldsymbol{\beta}^{*T}\right)^T$, το οποίο μεγιστοποιεί την αντίστοιχη λογαριθμοποιημένη συνάρτηση πιθανοφάνειας

$$l(\boldsymbol{\beta})=\sum_{i=1}^{n}\left\{\frac{y_i\bar{}\left[h\left(\ _0+\mathbf{x}_i^T\boldsymbol{\beta}^*\right)\right]+b\left(\bar{}\left[h\left(\ _0+\mathbf{x}_i^T\boldsymbol{\beta}^*\right)\right]\right)}{\ddot{}_i}+c\left(y_i,\ddot{}_i\right)\right\}. \quad (2)$$

Απλή παραγώγιση μας δίνει την ακόλουθη συνάρτηση (score function)

$$s(\boldsymbol{\beta}) = \frac{\partial l(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n} \frac{y_i - b'(\theta_i)}{Var(y_i)} \frac{\partial h(\eta_i)}{\partial \eta} \mathbf{x}_i = \mathbf{X}^T \mathbf{D} \boldsymbol{\Sigma}^{-1} (\mathbf{y} - \boldsymbol{\mu}), \quad (3)$$

όπου $\mathbf{X}^T = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$,

$$\mathbf{D} = diag\left\{ \frac{\partial h(\eta_1)}{\partial \eta}, \ldots, \frac{\partial h(\eta_n)}{\partial \eta} \right\}, \boldsymbol{\Sigma} = diag\left\{ Var(y_1), \ldots, Var(y_n) \right\}.$$

Ο πίνακας πληροφορίας του Fischer δίνεται από τον τύπο

$$F(\boldsymbol{\beta}) = -E\left[\frac{\partial^2 l(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} \right] = E\left[s(\boldsymbol{\beta}), s(\boldsymbol{\beta})^T \right] = \mathbf{X}^T \mathbf{W} \mathbf{X}, \quad (4)$$

όπου $\mathbf{W} = \mathbf{D} \boldsymbol{\Sigma}^{-1} \mathbf{D}^T$. Το άγνωστο διάνυσμα μπορεί να βρεθεί επαναληπτικά εφαρμόζοντας αριθμητικές μεθόδους για την επίλυση συστημάτων μη γραμμικών εξισώσεων, όπως η Newton-Raphson. Κάτω από ασθενείς υποθέσεις ο εκτιμητής μέγιστης πιθανοφάνειας $\hat{\boldsymbol{\beta}}$ είναι συνεπής και ασυμπτωτικά κανονικός με πίνακα διασποράς $Cov(\hat{\boldsymbol{\beta}}) = (\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1}$, βλέπε Fahrmeir & Kaufmann (1985).

Στην εργασία τους, οι Hoerl & Kennard (1970) έδειξαν οτι ο εκτιμητής ελαχίστων τετραγώνων του γραμμικού μοντέλου παλινδρόμησης τείνει να υπερεκτιμάει το μήκος του πραγματικού παραμετρικού διανύσματος, αν τα διανύσματα των μεταβλητών πρόβλεψης δεν είναι μεταξύ τους ορθογώνια. Ο Segerstedt (1992) έδειξε παρόμοια αποτελέσματα για τις εκτιμήσεις των γενικευμένων γραμμικών μοντέλων. Οι πρώτες προσπάθειες για γενικευμένη παλινδρόμηση κορυφογραμμής περιορίστηκαν στη λογιστική παλινδρόμηση, βλέπε για παράδειγμα Anderson & Blair (1982), Schaefer et al. (1984) και Duffy & Santner (1989). Ο Nyquist (1991) εισήγαγε την παλινδρόμηση κορυφογραμμής των γενικευμένων γραμμικών μοντέλων στο πλαίσιο της περιοριστικής εκτίμησης.

Όσο ο εκτιμητής μέγιστης πιθανοφάνειας της του άγνωστου παραμετρικού διανύσματος έχει την τάση να υπερεκτιμάει το μήκος του πραγματικού διανύσματος, είναι σκόπιμο να καθορίσουμε το τετραγωνικό του μήκος. Αυτός ο περιορισμός διαμορφώνεται ως εμπόδιο, και έτσι μπορούμε να χρησιμοποιήσουμε την Λαγκρατζιανή προσέγγιση. Τυπικά, λύνουμε το πρόβλημα βελτιστοποίησης

$$\hat{\boldsymbol{\beta}} = \arg\max_{\boldsymbol{\beta}} \left\{ l(\boldsymbol{\beta}) - P(\boldsymbol{\beta}) \right\}, \quad (5)$$

όπου

$$P(\boldsymbol{\beta}) = \alpha \|\boldsymbol{\beta}\|_2^2 = \alpha \sum_{j=1}^{p} \beta_j^2, \quad (6)$$

με $\|\boldsymbol{\beta}\|_2^2$ να ορίζεται η τετραγωνική L_2 νόρμα του $\boldsymbol{\beta}$ και $\alpha > 0$ είναι μία ρυθμιστική παράμετρος. Έστω $\hat{\boldsymbol{\beta}}_{ridge}(\alpha)$ να ορίζει τον προκύπτων GLM εκτιμητή παλινδρόμησης κορυφογραμμής για δοσμένο α. Συνεπώς, ο $\hat{\boldsymbol{\beta}}_{ridge}(\alpha)$ βασίζεται σε έναν L_2 όρο ποινής.

Συνήθως υπάρχει μία ρυθμιστική παράμετρος α, έτσι ώστε το ασυμπτωτικό μέσο τετραγωνικό σφάλμα του GLM εκτιμητή παλινδρόμησης κορυφογραμμής να είναι μικρότερο από την ασυμπτωτική διασπορά του εκτιμητή μέγιστης πιθανοφάνειας, για την απόδειξη βλέπε Segerstedt (1992). Παρ' όλα αυτά, το κύριο μειονέκτημα του $\hat{\boldsymbol{\beta}}_{ridge}(\alpha)$ είναι η έλλειψη του σε παραγωγή σποραδικών λύσεων.

Στις συνθήκες του γραμμικού μοντέλου, η πιο σημαντική ποινικοποιημένη μέθοδος η οποία αυτόματα περιλαμβάνει επιλογή καλύτερου υποσυνόλου είναι η Lasso, η οποία παρουσιάστηκε από τον Tibshirani (1996). Η L_1 ποινή της Lasso

$$P(\boldsymbol{\beta}) = \alpha \|\boldsymbol{\beta}\|_1 = \alpha \sum_{j=1}^{p} |\beta_j| \quad (7)$$

οδηγεί σε προσαρμοσμένα μοντέλα τα οποία είναι σποραδικά και ερμηνεύσιμα, με την έννοια οτι πολλές μεταβλητές δεν περιλαμβάνονται στο τελικό μοντέλο. Οι Shevade & Keerthi (2003) πρότειναν μία L_1 ποινικοποίηση για τη λογιστική παλινδρόμηση. Οι Park & Hastie (2007) εισήγαγαν έναν διορθωτικό-προβλεπτικό αλγόριθμο για γενικευμένα γραμμικά μοντέλα, που περιείχε Lasso ποινή. Το κύριο πρόβλημα με τη χρησιμοποίηση L_1 ποινών στο πλαίσιο των γενικευμένων γραμμικών μοντέλων είναι η αστάθεια των εκτιμηθέντων συντελεστών όταν κάποιες επεξηγηματικές μεταβλητές είναι ισχυρά συσχετισμένες. Επιπλέον, η λύση μπορεί να μην είναι μοναδική αν κάποιοι παλινδρομητές είναι πολυσυγγραμμικοί. Ως εκ τούτου, οι Park & Hastie (2007) τροποποίησαν τον όρο της Lasso ποινής σε

$$P(\boldsymbol{\beta}) = \alpha_1 \|\boldsymbol{\beta}\|_1 + \frac{\alpha_2}{2} \|\boldsymbol{\beta}\|_2^2, \quad (8)$$

όπου $\alpha_1 > 0$ είναι μία αυθαίρετη ρυθμιστική παράμετρος και α_2 είναι μία μικρή θετική σταθερά. Η Enet ποινή όπως παρουσιάστηκε από τους Zou & Hastie (2005) είναι αλγεβρικά ταυτόσημη με την (8), με μία αναβαθμονομημένη ρυθμιστική παράμετρο στον L_2 όρο ποινής. Η χρησιμοποίηση της (8) με

τον τρόπο των Zou & Hastie, απαιτεί ταυτόχρονη επιλογή των ρυθμιστικών παραμέτρων, για παράδειγμα με διασταυρωμένη επικύρωση, σε δύο διαστάσεις. Αυτό μπορεί να είναι υπολογιστικά περίπλοκο. Ένα κίνητρο που έδωσαν οι Zou & Hastie για την Enet, είναι η ιδιότητά της να περιλαμβάνει γκρουπ μεταβλητών που είναι ισχυρά συσχετισμένες. Όταν οι μεταβλητές είναι ισχυρά συσχετισμένες, όπως για παράδειγμα τα σετ δεδομένων με γονιδιακές εκφράσεις, η Lasso επιλέγει μόνο μία από το γκρουπ, ενώ η Enet επιλέγει όλο το γκρουπ.

Σε αυτό το κεφάλαιο παρουσιάζουμε μία εναλλακτική διαδικασία ποινικοποίησης η οποία σκοπεύει στην επιλογή των γκρουπ των συσχετισμένων μεταβλητών. Στην απλούστερη μορφή της βασίζεται σε μία ποινή η οποία σαφώς χρησιμοποιεί τη συσχέτιση μεταξύ των μεταβλητών ως βάρη. Στην επεκταμένη χρησιμοποιούνται ενισχυτικές τεχνικές για τα γκρουπ των μεταβλητών. Η ποινή ορίζεται ως

$$P_c(\boldsymbol{\beta}) = \alpha \sum_{i=1}^{p-1} \sum_{j>i} \left\{ \frac{(\beta_i - \beta_j)^2}{1-\rho_{ij}} + \frac{(\beta_i + \beta_j)^2}{1+\rho_{ij}} \right\} = 2\alpha \sum_{i=1}^{p-1} \sum_{j>i} \frac{\beta_i^2 - 2\rho_{ij}\beta_i\beta_j + \beta_j^2}{1-\rho_{ij}^2} \quad (9)$$

όπου ρ_{ij} δηλώνει τη συσχέτιση μεταξύ των i και j μεταβλητών. Έχει σχεδιαστεί ώστε να επικεντρώνεται στο αποτέλεσμα της ομαδοποίησης, έτσι ώστε ισχυρές συσχετίσεις να δίνουν συγκρίσιμες τιμές των εκτιμητών ($|\hat{\beta}_i| \approx |\hat{\beta}_j|$) με το πρόσημο να καθορίζεται από τη θετική ή την αρνητική συσχέτιση. Για ισχυρή θετική συσχέτιση ($\rho_{ij} \to 1$) ο πρώτος όρος γίνεται επικρατέστερος, έχοντας ως αποτέλεσμα οι εκτιμητές των β_i, β_j να είναι παρόμοιοι ($\hat{\beta}_i \approx \hat{\beta}_j$). Για ισχυρή αρνητική συσχέτιση ($\rho_{ij} \to -1$) ο δεύτερος όρος γίνεται επικρατέστερος και το $\hat{\beta}_i$ πλησιάζει το $-\hat{\beta}_j$. Συνεπώς, για αδύναμη συσχέτιση μεταξύ των δεδομένων το αποτέλεσμα είναι αρκετά κοντά στον εκτιμητή της παλινδρόμησης κορυφογραμμής. Η ποινή που βασίζεται στη συσχέτιση (9) μπορεί να γραφεί σε τετραγωνική μορφή

$$P_c(\boldsymbol{\beta}) = \alpha \boldsymbol{\beta}^T \mathbf{M} \boldsymbol{\beta}, \quad (10)$$

όπου ο πίνακας $\mathbf{M} = (m_{ij})$ δίνεται από τον τύπο

$$m_{ij} = \begin{cases} 2\sum_{s \neq i} \frac{1}{1-\rho_{is}^2}, & i = j, \\ -2\frac{\rho_{ij}}{1-\rho_{ij}^2}, & i \neq j. \end{cases}$$

Ορίζουμε τον προκύπτων ποινικοποιημένο εκτιμητή μέγιστης πιθανοφάνειας του άγνωστου διανύσματος των συντελεστών ως $\hat{\boldsymbol{\beta}}_c$ και στη συνέχεια θα αναφερόμαστε σε αυτόν ως GLM PenalReg εκτιμητή.

Εξαιτίας του συνδυασμού της λογαριθμοποιημένης συνάρτησης πιθανοφάνειας και της ποινής, ο υπολογισμός του GLM PenalReg εκτιμητή γίνεται εύκολα χρησιμοποιώντας τη score συνάρτηση και τον πίνακα του Fischer. Για την ποινικοποιημένη λογαριθμοποιημένη πιθανοφάνεια με $P_c\left(\boldsymbol{\beta}\right) = {}^{\mathsf{a}}\boldsymbol{\beta}^T\mathbf{M}\boldsymbol{\beta}$ λαμβάνουμε

$$l_p\left(\boldsymbol{\beta}\right) = l\left(\boldsymbol{\beta}\right) - \frac{\mathsf{a}}{2}\boldsymbol{\beta}^T\mathbf{M}\boldsymbol{\beta}, \quad (11)$$

όπου χρησιμοποιούμε μία ανακλιμάκωση της a για υπολογιστική απλοποίηση. Ως εκ τούτου, η ποινικοποιημένη score συνάρτηση είναι

$$s_p\left(\boldsymbol{\beta}\right) = \frac{\partial l_p\left(\boldsymbol{\beta}\right)}{\partial \boldsymbol{\beta}} = s\left(\boldsymbol{\beta}\right) - {}^{\mathsf{a}}\mathbf{M}\boldsymbol{\beta} = \mathbf{X}^T\mathbf{D}\boldsymbol{\Sigma}^{-1}\left(\mathbf{y}-\boldsymbol{\mu}\right) - {}^{\mathsf{a}}\mathbf{M}\boldsymbol{\beta}, \quad (12)$$

και ο ποινικοποιημένος πίνακας του Fischer είναι

$$F_p\left(\boldsymbol{\beta}\right) = -E\left[\frac{\partial s_p\left(\boldsymbol{\beta}\right)}{\partial \boldsymbol{\beta}^T}\right] = \mathbf{X}^T\mathbf{W}\mathbf{X} + {}^{\mathsf{a}}\mathbf{M}. \quad (13)$$

Όπως στον μη ποινικοποιημένο εκτιμητή μέγιστης πιθανοφάνειας χρειάζεται να επιλύσουμε ένα μη γραμμικό σύστημα εξισώσεων. Με τον ίδιο τρόπο όπως ο GLM εκτιμητής της παλινδρόμησης κορυφογραμμής, έτσι και ο GLM PenalReg εκτιμητής μπορεί να γραφεί ως ένας επαναληπτικός εκτιμητής ελαχίστων τετραγώνων με προσαρμοσμένα βάρη, ως έξης

$$\hat{\boldsymbol{\beta}}_c^{(k+1)} = \left(\mathbf{X}^T\mathbf{W}\mathbf{X} + {}^{\mathsf{a}}\mathbf{M}\right)^{-1}\mathbf{X}^T\mathbf{W}\tilde{\mathbf{y}}^{(k)}, \quad (14)$$

όπου $\tilde{\mathbf{y}}^{(k)} = \mathbf{X}\hat{\boldsymbol{\beta}}_c^{(k)} + \mathbf{D}^{-1}\left(\mathbf{y}-\boldsymbol{\mu}\right)$.

Βασισμένοι σε μία προσέγγιση Taylor πρώτης τάξης λαμβάνουμε τον πίνακα της ασυμπτωτικής διασποράς

$$Cov\left[\hat{\boldsymbol{\beta}}_c\left({}^{\mathsf{a}}\right)\right] = \left(\mathbf{X}^T\mathbf{W}\mathbf{X} + {}^{\mathsf{a}}\mathbf{M}\right)^{-1}\mathbf{X}^T\mathbf{W}\mathbf{X}\left(\mathbf{X}^T\mathbf{W}\mathbf{X} + {}^{\mathsf{a}}\mathbf{M}\right)^{-1}. \quad (15)$$

Παρατηρούμε οτι παίρνουμε παρόμοια αποτελέσματα με τον γενικευμένο εκτιμητή $\boldsymbol{\beta}_{ridge}\left(\mathrm{a}\right)$ όταν αντικαθιστούμε τον μοναδιαίο πίνακα με τον πίνακα ποινής $\mathbf{M}$, για λεπτομέρειες βλέπε Segerstedt (1992). Μία συστηματική αναφορά σε συγκρίσεις μέσων τετραγωνικών σφαλμάτων ανταγωνιστικών μεροληπτικών εκτιμητών για γραμμικά μοντέλα δίνεται από τους Trenkler & Toutenburg (1990). Στην ενότητα 3.4 αναφερόμαστε σε συγκρίσεις αποδοτικότητας μεθόδων πάνω σε διάφορα προσομοιωμένα και πραγματικά δεδομένα.

3.3 Γενικευμένη ενίσχυση κατά ομάδες

Το κύριο μειονέκτημα του εκτιμητή που βασίζεται στη συσχέτιση είναι η έλλειψη της σποραδικότητας. Πρακτικά, όταν θεωρούμε υψηλών διαστάσεων προβλήματα επιθυμούμε να επιλεγεί ένα κατάλληλο υποσύνολο μεταβλητών. Μία μέθοδος που είναι ικανή να ξεπεράσει αυτό το μειονέκτημα είναι η ενίσχυση κατά στοιχείο όπως παρουσιάστηκε από τους Bühlmann & Yu (2003). Πρότειναν την ανανέωση σε κάθε βήμα μόνο του στοιχείου βελτιώνει κατά μέγιστο τρόπο την προσαρμογή.

Οι ενισχυτικές μέθοδοι είναι πολλαπλά συστήματα πρόβλεψης τα οποία υπολογίζουν τις εκτιμηθείσες προβλέψεις από δεδομένα που έχει αναπροσαρμοστεί το βάρος τους. Προερχόμενη από την κοινότητα μηχανικής μάθησης το πρώτο κύριο πεδίο εφαρμογών της ήταν η δυαδική ταξινόμηση. Η σύνδεση μεταξύ της ενισχυτικής μεθόδου και μίας τεχνικής βελτιστοποίησης της κλίσης καθόδου της συνάρτησης όπως παρουσιάστηκε από τον Breiman (1998) παρείχε τη δυνατότητα εφαρμογής των ενισχυτικών μεθόδων και σε άλλα πλαίσια πέρα από την ταξινόμηση. Ο Friedman (2001) δημιούργησε τον L_2 Boost βασικό αλγόριθμο μάθησης, έναν αλγόριθμο βελτιστοποίησης με συνάρτηση απώλειας τετραγωνικού σφάλματος, ο οποίος μας παρέχει τα θεμέλια για την κατά στοιχείο ενίσχυση. Για μία λεπτομερή επισκόπηση πάνω σε αυτή τη διαδικασία βλέπε Meir & Rätsch (2003). Οι Tutz & Binder (2007) εφαρμόζουν την κατά στοιχείο ενίσχυση στον γενικευμένο εκτιμητή παλινδρόμησης κορυφογραμμής. Ο βασικός μαθητής αυτού του αλγόριθμου ενίσχυσης είναι το πρώτο βήμα του αλγόριθμου του Fischer.

Έστω $S^{(m)} \subset \{0,1,\ldots,p\}$ να ορίζει το σύνολο των δεικτών των μεταβλητών που θεωρούνται στο m βήμα, όπου ο δείκτης 0 αναφέρεται στο σταθερό όρο του διανύσματος. Τα δεδομένα εισαγωγής στο βασικό αλγόριθμο μάθησης είναι $\{(\mathbf{x}_1 r_1),\ldots,(\mathbf{x}_n r_n)\}$, όπου $r_i = y_i - \hat{\ll}_i^{(m-1)}, i = 1,\ldots,n$ ορίζεται το

υπόλοιπο μεταξύ της πραγματικής απόκρισης y_i και της εκτιμηθείσας απόκρισης από το προηγούμενο βήμα.

Η βασική ιδέα είναι να επιλέξουμε στο m βήμα της επαναληπτικής διαδικασίας εκείνο το υποσύνολο μεταβλητών που μας παρέχει την καλύτερη προσαρμογή. Στην κατά στοιχείο ενίσχυση που βασίζεται στη μέγιστη πιθανοφάνεια, συνήθως χρησιμοποιούμε την απόκλιση (deviance) ως μέτρο για την καταλληλότητα της προσαρμογής. Εμείς επιλέγουμε το κριτήριο πληροφορίας του Akaike (AIC) παρά την απόκλιση, επειδή αυτό περιλαμβάνει μία αυτόματη ποινικοποίηση των μεγάλων υποσυνόλων.

Ο ακόλουθος αλγόριθμος GenBlockBoost είναι μία έκδοση της ποινικοποιημένης εκτιμήτριας που βασίζεται στη συσχέτιση με ενισχυτικές τεχνικές.

<u>Αλγόριθμος GenBlockBoost</u>

Βήμα 1: (Αρχικοποίηση)

Προσαρμόζουμε το μοντέλο «$_i = h(\ _0)$ από τον επαναληπτικό αλγόριθμο του Fischer και παίρνουμε το $\hat{\boldsymbol{\beta}}^{(0)} = (\ \hat{}_0, 0, \ldots, 0)^T$. Θέτουμε $\hat{\boldsymbol{\eta}}^{(0)} = X\hat{\boldsymbol{\beta}}^{(0)}, \hat{\boldsymbol{\mu}}^{(0)} = h\left(\hat{\boldsymbol{\eta}}^{(0)}\right)$.

Βήμα 2: (Επανάληψη)

Για $m = 1, 2, \ldots$

a) Κατατάσσουμε κατάλληλα τις μεταβλητές σύμφωνα με το πόσο βελτιώνουν την προσαρμογή του μοντέλου

Για $j \in \{0, \ldots, p\}$ υπολογίζουμε τις εκτιμήτριες με μία επανάληψη του αλγόριθμου του Fischer

$$\hat{b}_{\{j\}} = \left(\mathbf{x}_{\{j\}}^T W\left(\hat{\boldsymbol{\eta}}^{(m-1)}\right)\mathbf{x}_{\{j\}} + \text{a}\right)^{-1} \mathbf{x}_{\{j\}}^T W\left(\hat{\boldsymbol{\eta}}^{(m-1)}\right) D\left(\hat{\boldsymbol{\eta}}^{(m-1)}\right)^{-1} \left(\mathbf{y} - \hat{\boldsymbol{\mu}}^{(m-1)}\right),$$

κατατάσσοντας τα $\hat{b}_{j_0}, \ldots, \hat{b}_{j_p}$ έτσι ώστε $Dev\left(\hat{b}_{j_0}\right) \le \ldots \le Dev\left(\hat{b}_{j_p}\right)$, όπου

$$Dev\left(\hat{b}_{j_k}\right) = 2\sum_{i=1}^{n}\left\{l_i\left(y_i\right) - l_i\left[h\left(\text{r}_i^{(m-1)} + x_{ij_k}\hat{b}_{j_k}\right)\right]\right\}, k = 0, 1, \ldots, p\,.$$

b) Βρίσκουμε τον κατάλληλο αριθμό μεταβλητών για ανανέωση

Για $r = 0, \ldots, p$

Με $S_r = \{ j_0, \ldots, j_r \}$ υπολογίζουμε τις εκτιμήτριες με μία επανάληψη του αλγόριθμου του Fischer

$$\hat{\mathbf{b}}_{S_r} = \left(\mathbf{X}_{S_r}^T W\left(\hat{\boldsymbol{\eta}}^{(m-1)}\right) \mathbf{X}_{S_r} + \mathsf{a}_{|S_r|} \mathbf{M}_{S_r} \right)^{-1} \mathbf{X}_{S_r}^T W\left(\hat{\boldsymbol{\eta}}^{(m-1)}\right) \times D\left(\hat{\boldsymbol{\eta}}^{(m-1)}\right)^{-1} \left(\mathbf{y} - \hat{\boldsymbol{\mu}}^{(m-1)}\right),$$

παίρνοντας τις εκτιμήτριες $\hat{\mathbf{b}}_{S_r}$ και τα κριτήρια πληροφορίας $AIC\left(\hat{\mathbf{b}}_{S_r}\right)$.

c) Επιλογή

Επιλέγουμε το υποσύνολο των μεταβλητών που έχει την καλύτερη προσαρμογή

$$S^{(m)} = \arg\min_{S_r} AIC(\hat{\mathbf{b}}_{S_r}).$$

d) Αναπροσαρμογή

Το διάνυσμα των παραμέτρων ανανεώνεται ως εξής:

$$\hat{}_j^{(m)} = \begin{cases} \hat{}_j^{(m-1)} + \hat{b}_j, & \text{αν j} \in \mathrm{S}^{(m)}, \\ \hat{}_j^{(m-1)}, & \text{αλλιώς}, \end{cases}$$

παίρνοντας $\hat{\boldsymbol{\beta}}^{(m)} = \left(\hat{}_1^{(m)}, \ldots, \hat{}_p^{(m)} \right)^T$, $\hat{\boldsymbol{\eta}}^{(m)} = \mathbf{X}\hat{\boldsymbol{\beta}}^{(m)}$, $\hat{\boldsymbol{\mu}}^{(m)} = h\left(\hat{\boldsymbol{\eta}}^{(m)}\right)$.

Ο αριθμός των δυνατών συνδυασμών των μεταβλητών είναι 2^p. Εξαιτίας υπολογιστικών περιορισμών δεν μπορούμε να κάνουμε ένα πλήρες ψάξιμο για το καλύτερο υποσύνολο. Ως εκ τούτου στο πρώτο βήμα κάθε επανάληψης κατατάσσουμε τις μεταβλητές σύμφωνα με τη ατομική τους δυνατότητα στη βελτίωση της προσαρμογής. Αυτή η βελτίωση μετριέται με την δυνητική τους απόκλιση

$$Dev\left(\hat{b}_j\right) = 2\sum_{i=1}^{n} \left\{ l_i\left(y_i\right) - l_i\left[h\left(\hat{}_i^{(m-1)} + x_{ij}\hat{b}_j \right) \right] \right\}, \qquad j = 0, \ldots, p,$$

όπου $x_{i0} = 1$ για όλα τα $i = 1, \ldots, n$.

Για να κάνουμε τον βασικό αλγόριθμο μάθησης αδύναμο, έτσι ώστε να γίνονται μόνο μικρές αλλαγές στην εκτίμηση της παραμέτρου σε κάθε επανάληψη, η ρυθμιστική παράμετρος a επιλέγεται πολύ μεγάλη. Αυτό επίσης οδηγεί σε πιο ευσταθείς εκτιμήτριες. Το αντίτιμο γι' αυτήν την επιλογή είναι η αύξηση του χρόνου υπολογισμού όσο αυξάνεται η τιμή της παραμέτρου.

Για τα υποσύνολα S που περιέχουν μία μόνο μεταβλητή η ποινή που βασίζεται στη συσχέτιση (10) δεν μπορεί να χρησιμοποιηθεί άμεσα. Σε αυτές τις περιπτώσεις εφαρμόζουμε τη ποινή της παλινδρόμησης κορυφογραμμής $P_{c,\{j\}} = \mathsf{a}_j^2$.

Μέσα στον αλγόριθμο ο εκτιμητής που βασίζεται στη συσχέτιση των μεταβλητών χρησιμοποιείται για υποσύνολα διαφόρων μεγεθών. Η ρυθμιστική παράμετρος a που χρησιμοποιείται πρέπει να προσαρμόζεται στον αριθμό των μεταβλητών που ανανεώνονται. Αν θεωρήσουμε την περίπτωση των ασυσχέτιστων μεταβλητών η ποινή γίνεται $P_c\left(\boldsymbol{\beta}\right)=2^{\mathsf{a}}\left(p-1\right)\sum_{j=1}^{p}{}^{2}_{j}$ η οποία ισούται με τη ποινή της παλινδρόμησης κορυφογραμμής με ρυθμιστική παράμετρο $2^{\mathsf{a}}\left(p-1\right)$. Γι' αυτό η $^{\mathsf{a}}{}_{|S_r|}$ στο βήμα 2b του GenBlockBoost επιλέγεται να είναι $^{\mathsf{a}}{}_{|S_r|}=^{\mathsf{a}}\left(\left|S_r\right|-1\right)$, όπου $\left|S_r\right|$ ορίζεται η πληθικότητα του συνόλου των μεταβλητών που ανανεώνονται.

Με σκοπό να αποφύγουμε την υπερπροσαρμογή, χρειάζεται ένα κριτήριο τερματισμού που να εκτιμάει τον βέλτιστο αριθμό επαναλήψεων. Χρησιμοποιούμε το AIC κριτήριο

$$AIC\left(\hat{\boldsymbol{\beta}}^{(m)}\right)=Dev_m+2tr\left(\mathbf{H}_m\right), \quad (16)$$

με

$$Dev_m=2\sum_{i=1}^{n}\left[l_i\left(y_i\right)-l_i\left(\hat{\ll}_i^{(m)}\right)\right].$$

Μία προσέγγιση του πίνακα προβολής είναι

$$\mathbf{H}_m=\sum_{j=0}^{m}\mathbf{M}_j\prod_{i=0}^{j-1}\left(I-\mathbf{M}_i\right),$$

έτσι ώστε $\hat{\boldsymbol{\mu}}^{(m)}=\mathbf{H}_m\mathbf{y}$, όπου

$$\mathbf{M}_l=\boldsymbol{\Sigma}_m^{1/2}\mathbf{W}_m^{1/2}\mathbf{X}_{S^{(m)}}\left(\mathbf{X}_{S^{(m)}}^T\mathbf{W}_m\mathbf{X}_{S^{(m)}}+{}^{\mathsf{a}}\mathbf{M}_{S^{(m)}}\right)^{-1}\mathbf{X}_{S^{(m)}}^T\mathbf{W}_m^{1/2}\boldsymbol{\Sigma}_m^{-1/2}$$

και $\mathbf{M}_0=\frac{1}{n}1_n1_n^T$. Βλέπε Tutz & Leitenstorfer (2007) για το πώς προήλθε αυτή η προσέγγιση. Μία εκτίμηση για τον κατάλληλο αριθμό επαναλήψεων είναι

$$m^*=\arg\min_m AIC\left(\hat{\boldsymbol{\beta}}^{(m)}\right).$$

Στην επόμενη ενότητα ερευνούμε την αποδοτικότητα του ποινικοποιημένου εκτιμητή που βασίζεται στη συσχέτιση για τα γενικευμένα γραμμικά μοντέλα και του αλγόριθμου GenBlockBoost πάνω σε διάφορες προσομοιώσεις και σετ δεδομένων.

3.4 Προσομοιώσεις και παραδείγματα πραγματικών δεδομένων (λευχαιμία)

Στις προσομοιώσεις θεωρούμε μεταβλητές οι οποίες δίνονται σε 10 γκρουπ, κάθε γκρουπ περιέχει q μεταβλητές, καταλήγοντας σε $p=10q$ μεταβλητές συνολικά. Όλες οι μεταβλητές έχουν μοναδιαία διασπορά. Οι συσχετίσεις μεταξύ x_i και x_j είναι $\circ^{|i-j|}$ αν x_i και x_j ανήκουν στο ίδιο γκρουπ, διαφορετικά δίνονται από μία αποκομμένη κανονική κατανομή $N\left(0,0.1^2\right)$. Για τη γραμμική παράμετρο πρόβλεψης ¦ επιλέγουμε τα σύνολα V των μεταβλητών που ανήκουν σε τρία τυχαία επιλεγμένα γκρουπ έτσι ώστε

$$¦ = \mathbf{x}^T\boldsymbol{\beta},$$

όπου $\mathbf{x}=\left(x_1,\ldots,x_p\right)^T$ και το $\boldsymbol{\beta}=c\cdot\left(_1,\ldots, _p\right)^T$ καθορίζεται από

$$_j \sim N(1,1) \text{ για } j\in V, \quad _j=0 \text{ αλλιώς.}$$

Αυτό σημαίνει ότι κάθε μεταβλητή που περιλαμβάνεται σε ένα απ' τα επιλεγμένα γκρουπ θεωρείται σημαντική. Σημειώνουμε οτι $_0=0$ σε όλες τις προσομοιώσεις, αλλά όλες επιτρέπεται να περιλαμβάνουν μία μη μηδενική σταθερά στο διάνυσμα των εκτιμηθέντων συντελεστών. Η τελική απόκριση y αντιστοιχεί στην αναμενόμενη τιμή της απόκρισης $« = E\left(y|\mathbf{x}\right)=h(¦)$, όπου η $h(¦)=\exp(¦)/\left(1+\exp(¦)\right)$ σχεδιάζεται από μία διωνυμική κατανομή $B(«,1)$. Η σταθερά c επιλέγεται έτσι ώστε ο λόγος

$$ratio=\frac{\sum_{i=1}^{n}\left(«_i-\bar{«}\right)^2}{\sum_{i=1}^{n}Var\left(y_i\right)},$$

με $\bar{«}=\frac{1}{n}\sum_{i=1}^{n}«_i$, να είναι προσεγγιστικά ίσος με τη μονάδα. Χρησιμοποιούμε τον αλγόριθμο του Newton για να βρούμε το c. Η εκτίμηση των άγνωστων παραμέτρων βασίζεται σε σύνολο εκπαίδευσης 100 παρατηρήσεων. Για την επικύρωση χρησιμοποιούμε σύνολο ελέγχου 1000 παρατηρήσεων. Τέλος, χρησιμοποιούμε ένα πρόσθετο ανεξάρτητο σύνολο επικύρωσης αποτελούμενο από 100 παρατηρήσεις για τον προσδιορισμό των ρυθμιστικών παραμέτρων.

Συγκρίνουμε τον GLM PenalReg εκτιμητή και τον GenBlockBoost αλγόριθμο με τον εκτιμητή μέγιστης πιθανοφάνειας (ML), τον L_2 ποινικοποιημένο εκτιμητή μέγιστης πιθανοφάνειας (παλινδρόμηση κορυφογραμμής), τον L_1 ποινικοποιημένο εκτιμητή μέγιστης πιθανοφάνειας (Lasso) και μία ενισχυτική έκδοση του L_2 εκτιμητή (GenRidgeBoost). Για περισσότερες λεπτομέρειες πάνω στον GenRidgeBoost βλέπε Tutz & Binder (2007). Ο υπολογισμός του L_1 εκτιμητή έγινε με το πακέτο glmpath της R από τους Mee Young Park & Trevor Hastie.

Η αποδοτικότητα των μεθόδων μετριέται από τα MSE_y και MSE . Το τελευταίο ορίζεται ως

$$MSE = \left|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right|^2 . \quad (17)$$

Εκτός από την προβλεπτική ικανότητα, ένα επίσης σημαντικό κριτήριο για τη σύγκριση των μεθόδων είναι ο αριθμός των μεταβλητών που περιλαμβάνονται στο τελικό μοντέλο. Το τελικό μοντέλο πρέπει να είναι όσο πιο οικονομικό γίνεται αλλά όλες οι σημαντικές μεταβλητές πρέπει να περιλαμβάνονται. Χρησιμοποιούμε τα κριτήρια ‘επιτυχίες’ και ‘αποτυχίες’ για να αξιολογήσουμε τον προσδιορισμό των σημαντικών μεταβλητών. Οι ‘επιτυχίες’ αναφέρονται στον αριθμό των σωστά προσδιορισμένων σημαντικών μεταβλητών, οι ‘αποτυχίες’ στον αριθμό των μη σημαντικών που η μέθοδος τις αξιολόγησε σημαντικές.

Τα αποτελέσματα των προσομοιώσεων δίνονται στους Πίνακες 1,2,3 και στα Γραφήματα 1 και 2. Ο αλγόριθμος GenBlockBoost έχει την καλύτερη προβλεπτικότητα σχεδόν κάθε φορά. Θεωρώντας την προσαρμογή των πραγματικών παραμέτρων ο PenalReg εκτιμητής λειτουργεί πολύ καλά, αλλά ο GenBlockBoost παρουσιάζει καλά αποτελέσματα στην επιλογή μεταβλητών σε μικρά και μεσαίου μεγέθους γκρουπ. Ο glmpath λειτουργεί καλύτερα σε μεγάλα γκρουπ. Όσον αφορά τις ‘επιτυχίες’ και ‘αποτυχίες’ ο GenBlockBoost υπερέχει καθαρά του glmpath και επίσης επιλέγει περισσότερες σημαντικές μεταβλητές από τον GenRidgeBoost. Ο GenRidgeBoost γενικά τείνει σε πιο οικονομικά μοντέλα, συνεπώς η διάμεσος των ‘αποτυχιών’ είναι μικρότερη σε σύγκριση με αυτήν του GenBlockBoost.

Για εφαρμογή πάνω σε πραγματικά δεδομένα χρησιμοποιούμε το σετ δεδομένων της λευχαιμίας όπως περιγράφτηκε από τους Golub et al. (1999). Για τη θεραπεία του καρκίνου είναι σημαντικό να στοχεύσεις συγκεκριμένες θεραπείες σε παθογενετικά διακριτούς τύπους όγκου, για να κερδίσεις το μέγιστο της αποτελεσματικότητας και το ελάχιστο της τοξικότητας. Ως εκ τούτου, ο διαχωρισμός των διαφορετικών τύπων όγκου είναι κομβικής σημασίας για επιτυχημένη θεραπεία. Η πρόκληση στο σετ δεδομένων της λευχαιμίας είναι να ταξινομήσεις την οξεία λευχαιμία σε αυτές που προκύπτουν από λεμφοειδείς προδρόμους (οξεία λεμφοβλαστική λευχαιμία, ALL) και αυτές που προκύπτουν από

μυελοειδείς προδρόμους (οξεία μυελογενής λευχαιμία, AML), βασισμένο στην ταυτόχρονη παρακολούθηση 7129 γονιδιακών εκφράσεων. Το σετ δεδομένων αποτελείται από 72 δείγματα, εκ των οποίων 47 παρατηρήσεις είναι ALL και 25 είναι AML. Χρησιμοποιούμε 20 τυχαία 'σπασίματα' του σετ σε ένα σύνολο εκπαίδευσης και ένα σύνολο ελέγχου, 38 και 34 παρατηρήσεων αντίστοιχα.

Εκτός από το τεστ απόκλισης

$$Dev_{test} = 2\sum_{i=1}^{n_{test}}\left[l_i\left(y_{i,test}\right) - l_i\left(\hat{«}_{i,test}\right)\right], \quad (18)$$

που βασίζεται στο σύνολο ελέγχου, θεωρούμε τον αριθμό των γονιδίων που προσδιορίστηκαν ως σημαντικές μεταβλητές. Όσο ο κύριος στόχος είναι η ταξινόμηση, επικεντρωνόμαστε στον αριθμό των σωστά ταξινομημένων και αντίστοιχα λανθασμένα ταξινομημένων παρατηρήσεων πάνω στο σύνολο ελέγχου, ως μέτρα αποδοτικότητας.

Εξ αιτίας των 20 τυχαίων 'σπασιμάτων' θεωρούμε τα μέσα αποτελέσματα απόδοσης, που δίνονται στον Πίνακα 4. Και οι τρεις αλγόριθμοι παρουσιάζουν σχεδόν παρόμοια απόδοση. Στο μέσο αριθμό των σωστά ταξινομημένων τύπων λευχαιμίας, ο GenRidgeBoost είναι ελαφρώς καλύτερος για την ALL ομάδα, ο GenBlockBoost ελαφρώς καλύτερος για την AML ομάδα. Όταν θεωρούμε την ολική λανθασμένη ταξινόμηση ο glmpath έχει το καλύτερο αποτέλεσμα. Λόγω του τεστ απόκλισης, το σετ δεδομένων προσαρμόζεται καλύτερα στο μοντέλο που προκύπτει από τον GenBlockBoost. Εδώ, ο GenRidgeBoost εκτιμητής είναι φτωχός. Όταν θεωρούμε τον αριθμό των επιλεγμένων γονιδίων ο GenRidgeBoost είναι ελαφρώς πιο σποραδικός από τους υπόλοιπους.

		ML	Ridge	PenalReg	GenRidgeBoost	GenBlockBoost	GlmPath (Lasso)
$q = 3$	$\varrho = 0.95$	17983.21	875.22	**866.16**	965.81	901.05	904.78
	$\varrho = 0.8$	16791.35	928.75	923.54	916.89	**907.23**	940.89
	$\varrho = 0.5$	15497.62	965.58	966.91	890.67	**881.59**	936.87
$q = 5$	$\varrho = 0.95$	20035.41	894.60	892.68	891.95	**851.78**	908.84
	$\varrho = 0.8$	21152.08	939.29	934.00	906.33	**897.05**	949.74
	$\varrho = 0.5$	19842.48	1005.99	1011.70	993.47	**958.38**	1007.23
$q = 10$	$\varrho = 0.95$	-	871.39	**854.36**	868.69	859.35	907.84
	$\varrho = 0.8$	-	970.10	947.54	937.15	**915.58**	982.49
	$\varrho = 0.5$	-	1099.91	1085.18	1119.54	1110.80	**1083.11**

Πίνακας 6: *Μέσες αποκλίσεις των προσομοιώσεων έπειτα από 20 επαναλήψεις.*

		ML	Ridge	PenalReg	GenRidgeBoost	GenBlockBoost	GlmPath (Lasso)
$q = 3$	$\varrho = 0.95$	423640.00	2.19	**1.80**	2.69	2.56	3.55
	$\varrho = 0.8$	106086.80	1.98	1.68	1.89	**1.62**	1.92
	$\varrho = 0.5$	47861.17	2.00	2.04	**1.30**	1.43	1.63
$q = 5$	$\varrho = 0.95$	345348.10	1.60	**1.51**	3.62	2.07	3.71
	$\varrho = 0.8$	77118.91	2.35	1.97	2.27	**1.95**	2.80
	$\varrho = 0.5$	33738.83	2.15	2.19	2.12	**1.78**	2.18
$q = 10$	$\varrho = 0.95$	-	1.43	**1.08**	2.87	2.40	2.22
	$\varrho = 0.8$	-	2.02	1.55	2.72	**2.49**	2.46
	$\varrho = 0.5$	-	2.51	**2.38**	2.67	2.79	2.58

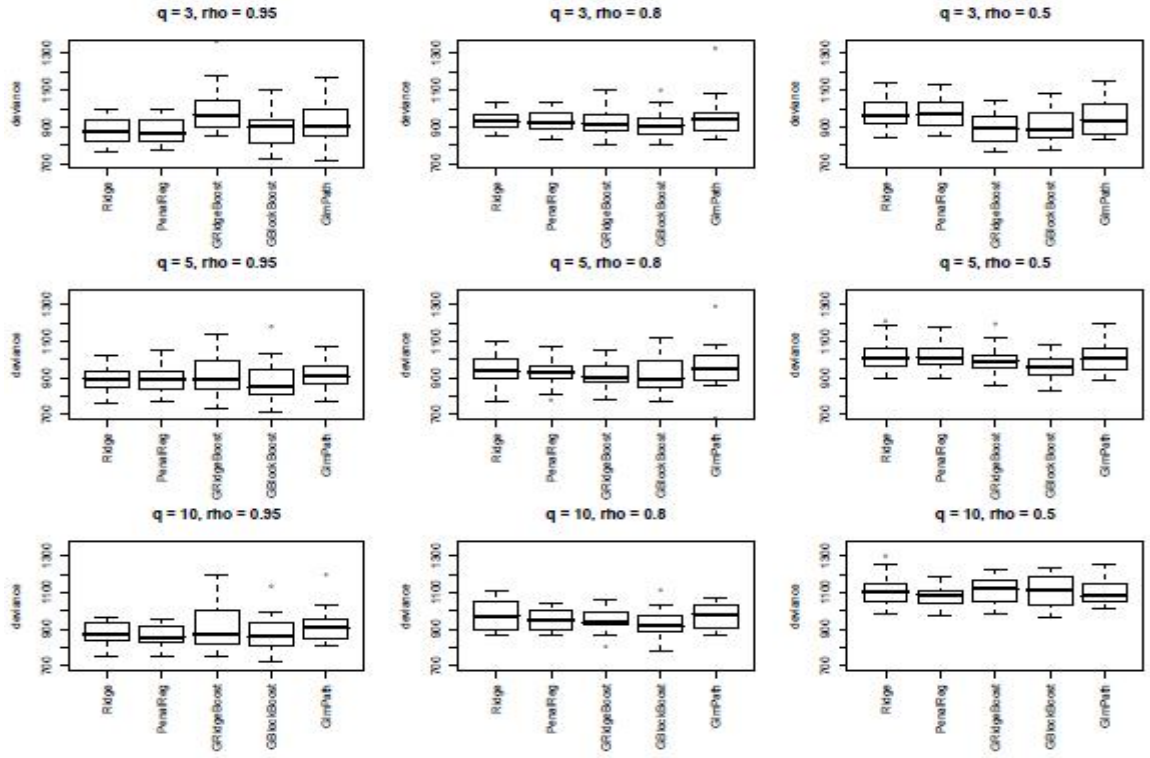
Πίνακας 7: *Μέσα MSE των προσομοιώσεων έπειτα από 20 επαναλήψεις.*

		ML	Ridge	PenalReg	GenRidgeBoost	GenBlockBoost	GlmPath (Lasso)
$q = 3$	$\varrho = 0.95$	9/22	9/22	9/22	4/1	6/3	5/7
	$\varrho = 0.8$	9/22	9/22	9/22	5/1	5/2	6/8
	$\varrho = 0.5$	9/22	9/22	9/22	6/2	6/3	7/10
$q = 5$	$\varrho = 0.95$	15/36	15/36	15/36	6/3	12/4	6/9
	$\varrho = 0.8$	15/36	15/36	15/36	7/3	11/6	8/9
	$\varrho = 0.5$	15/36	15/36	15/36	8/3	9/5	9/9
$q = 10$	$\varrho = 0.95$	-	30/71	30/71	9/2	17/8	8/5
	$\varrho = 0.8$	-	30/71	30/71	10/2	16/5	12/10
	$\varrho = 0.5$	-	30/71	30/71	12/2	17/9	14/12

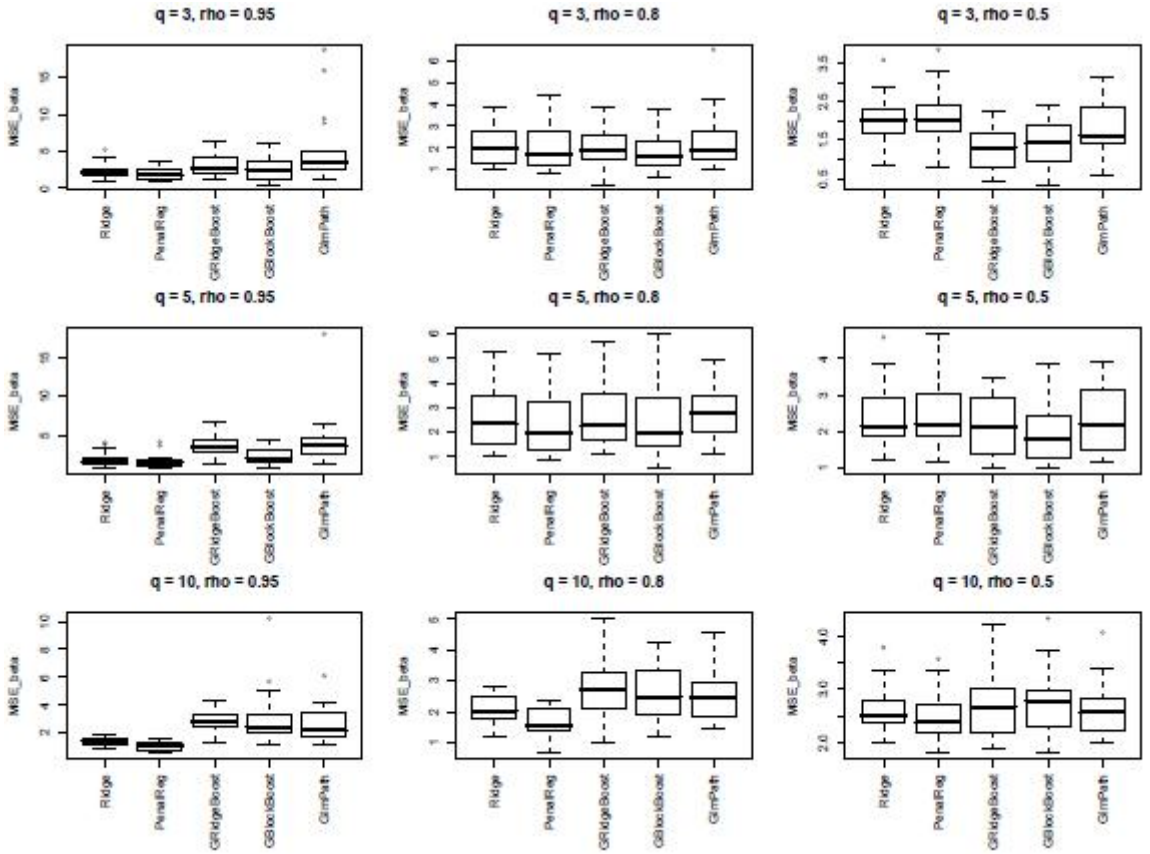
Πίνακας 8: *Μέσες 'επιτυχίες'/ 'αποτυχίες' των προσομοιωμένων δεδομένων έπειτα από 20 επαναλήψεις.*

Performance measure	GenBlockBoost	GlmPath	GenRidgeBoost
ALL correctly classified	9	10	11
AML correctly classified	21	20	20
misclassification	5	3	5
Dev_{test}	17.75	19.11	84.98
No. of genes used	11	10	9

Πίνακας 9: *Μέσα αποτελέσματα για το σετ δεδομένων της λευχαιμίας με 20 τυχαία 'σπασίματα'.*



Γράφημα 10: *Αποκλίσεις των διάφορων εκτιμητών.*



Γράφημα 11: *Τα* MSE *των διάφορων εκτιμητών.*

3.5 Συμπεράσματα

Παρουσιάσαμε δύο διαδικασίες για την εκτίμηση παραμέτρων σε γενικευμένα γραμμικά μοντέλα με πολλές μεταβλητές. Ο GLM PenalReg εκτιμητής δίνει ιδιαίτερη προσοχή στο αποτέλεσμα της ομαδοποίησης, ο αλγόριθμος GenBlockBoost δίνει επιπλέον επιλογή υποσυνόλου μεταβλητών. Οι προσομοιώσεις δείχνουν της ανταγωνιστική απόδοση στην προσαρμογή των δεδομένων και τη μικρή απόκλιση μεταξύ εκτιμηθέντων και πραγματικών παραμέτρων. Ο GenBlockBoost είναι ελαφρώς λιγότερο σποραδικός από τον GenRidgeBoost αλλά αυτό είναι συνέπεια του οτι επικεντρώνεται πιο σφιχτά στο αποτέλεσμα της ομαδοποίησης. Παρ' όλα αυτά ο σωστός προσδιορισμός των σημαντικών μεταβλητών είναι αρκετά καλός. Ως συμπέρασμα, ο GenBlockBoost εκτιμητής μπορεί να χαρακτηριστεί ως ένας σκληρός ανταγωνιστής στο πεδίο της επιλογής μεταβλητών σε γενικευμένα γραμμικά μοντέλα.

Και οι δύο μέθοδοι μπορούν να επεκταθούν στην περίπτωση των πολυμεταβλητών γενικευμένων γραμμικών μοντέλων, όπως αυτά με πολυωνυμική απόκριση. Επιπλέον, κάποιες εξτρά θεωρητικές πλευρές πάνω στις συγκρίσεις των MSE με τον GLM εκτιμητή της παλινδρόμησης κορυφογραμμής θα ήταν ενδιαφέρουσες.

ΠΑΡΑΡΤΗΜΑ

Απόδειξη του Θεωρήματος 1.3.3.1.1

Έστω ότι $\sim_n = n^{-1/2} + a_n$. Θέλουμε να δείξουμε ότι για οποιοδήποτε $£ > 0$, υπάρχει σταθερά C, ώστε

$$P\left\{\sup_{\|\underset{\sim}{u}\|=C} Q(\underset{\sim}{}_0 + \sim_n \underset{\sim}{u}) < Q(\underset{\sim}{}_0)\right\} \geq 1 - £ \quad (5.1).$$

Αυτό συνεπάγεται, ότι με πιθανότητα τουλάχιστον $1-£$, θα υπάρχει ένα τοπικό μέγιστο που θα ανήκει στη σφαίρα $\left\{\underset{\sim}{}_0 + \sim_n \underset{\sim}{u} : \| \underset{\sim}{u} \| \leq C\right\}$. Συνεπώς, θα υπάρχει ένα τοπικό ελάχιστο τέτοιο ώστε $\| \hat{\underset{\sim}{}} - \underset{\sim}{}_0 \| = O_p(\sim_n)$. Χρησιμοποιώντας τώρα $p_{a_n}(0) = 0$, έχουμε

$$D_n(\underset{\sim}{u}) \equiv Q(\underset{\sim}{}_0 + \sim_n \underset{\sim}{u}) - Q(\underset{\sim}{}_0) \leq L(\underset{\sim}{}_0 + \sim_n \underset{\sim}{u}) - L(\underset{\sim}{}_0) - n\sum_{j=1}^{s}\left\{p_{a_n}(|\quad_{j0} + \sim_n u_j|) - p_{a_n}(|\quad_{j0})\right\}$$ όπου s είναι ο αριθμός των συνιστωσών του $\underset{\sim}{}_{10}$. Έστω τώρα $L'(\underset{\sim}{}_0)$ να είναι το βαθμωτό διάνυσμα του L. Χρησιμοποιώντας επέκταση Taylor της συνάρτησης πιθανοφάνειας, έχουμε

$$D_n(\underset{\sim}{u}) \leq \sim_n L'(\underset{\sim}{}_0)'\underset{\sim}{u} - \frac{1}{2}\underset{\sim}{u}' I(\underset{\sim}{}_0)\underset{\sim}{u} n\sim_n^2 \{1 + o_p(1)\}$$

$$-\sum_{j=1}^{s}\left\{n\sim_n p'_{a_n}(|\quad_{j0}|\,\mathrm{sgn}(|\quad_{j0}|)u_j + n\sim_n^2 p''_{a_n}(|\quad_{j0}|)u_j^2\{1 + o_p(1)\}\right\} \quad (5.2).$$

Να παρατηρήσουμε ότι $n^{-1/2}L'(\underset{\sim}{}_0) = O_p(1)$. Συνεπώς, ο πρώτος όρος του δεξιού μέλους της (5.2) είναι της τάξης $O_p(n^{1/2}\sim_n) = O_p(n\sim_n^2)$. Επιλέγοντας αρκετά μεγάλο C, ο δεύτερος όρος επικρατεί του πρώτου ομοιόμορφα στο $\| \underset{\sim}{u} \| = C$. Επίσης, ο τρίτος όρος φράσσεται από την ποσότητα

$$\sqrt{s}n\sim_n a_n \| \underset{\sim}{u} \| + n\sim_n^2 \max\left\{|\, p''_{a_n}(|\quad_{j0}|)| : \quad_{j0} \neq 0\right\} \| \underset{\sim}{u} \|^2.$$

Ο δεύτερος όρος της (5.2) επικρατεί και εδώ. Οπότε, με την επιλογή επαρκώς μεγάλου C, η (5.1) ισχύει και αυτό ολοκληρώνει την απόδειξη.

Απόδειξη του Λήμματος 1.3.3.1.1

Πρέπει να δείξουμε ότι, με πιθανότητα να τείνει στο 1 όσο το $n \to \infty$, για οποιοδήποτε $\tilde{}_1$ που ικανοποιεί $\tilde{}_1 - \tilde{}_{10} = O_p(n^{-1/2})$ και για κάποιο $£_n = Cn^{-1/2}$ και $j = s+1,\ldots,d$, ισχύει

$$\frac{\partial Q(\tilde{})}{\partial \ _j} < 0, \text{ για } 0 < \ _j < £_n \quad (5.3)$$

$$\frac{\partial Q(\tilde{})}{\partial \ _j} > 0, \text{ για } -£_n < \ _j < 0 \quad (5.4).$$

Για να δείξουμε ότι ισχύει η (5.3), χρησιμοποιώντας επέκταση Taylor, έχουμε

$$\frac{\partial Q(\tilde{})}{\partial \ _j} = \frac{\partial L(\tilde{})}{\partial \ _j} - np'_{a_n}(|\ _j|)\operatorname{sgn}(|\ _j|)$$

$$= \frac{\partial L(\tilde{}_0)}{\partial \ _j} + \sum_{l=1}^{d} \frac{\partial^2 L(\tilde{}_0)}{\partial \ _j \partial \ _l}(\ _l - \ _{l0})$$

$$+ \sum_{l=1}^{d}\sum_{k=1}^{d} \frac{\partial^3 L(\tilde{}^*)}{\partial \ _j \partial \ _l \partial \ _k} \times (\ _l - \ _{l0})(\ _k - \ _{k0}) - np'_{a_n}(|\ _j|)\operatorname{sgn}(|\ _j|),$$

όπου το $\tilde{}^*$ είναι μεταξύ των $\tilde{}$ και $\tilde{}_0$. Να σημειώσουμε ότι

$$n^{-1}\frac{\partial L(\tilde{}_0)}{\partial \ _j} = O_p(n^{-1/2})$$

και

$$\frac{1}{n}\frac{\partial^2 L(\tilde{}_0)}{\partial \ _j \partial \ _l} = E\left\{\frac{\partial^2 L(\tilde{}_0)}{\partial \ _j \partial \ _l}\right\} + o_p(1).$$

Από την υπόθεση ότι $\tilde{} - \tilde{}_0 = O_p(n^{-1/2})$, έχουμε τώρα ότι

$$\frac{\partial Q(\tilde{})}{\partial \ _j} = n\mathsf{a}_n\left\{-\mathsf{a}_n^{-1}p'_{a_n}(|\ _j|)\operatorname{sgn}(|\ _j|) + O_p(n^{-1/2}/\mathsf{a}_n\right\}.$$

Λαμβάνοντας υπόψην ότι

$\liminf_{n\to\infty} \liminf_{\theta\to 0^+} a_n^{-1} p'_{a_n}(\theta) > 0$ και $n^{-1/2}/a_n \to 0$,

το πρόσημο της παραγώγου εξαρτάται αποκλειστικά από αυτό του β_j. Συνεπώς, οι (5.3) και (5.4) ισχύουν.

Απόδειξη του Θεωρήματος 1.3.3.1.2

Από το προηγούμενο Λήμμα, έχουμε ότι ισχύει η σποραδικότητα. Θα αποδείξουμε τώρα την ασυμπτωτική κανονικότητα. Μπορεί εύκολα να δειχθεί ότι υπάρχει ένα $\hat{\beta}_1$ στο πρώτο Θεώρημα το οποίο είναι ένα $\sqrt{n}$ -συνεπές τοπικό μέγιστο του $Q\left\{\begin{pmatrix}\beta_1\\ 0\end{pmatrix}\right\}$, το οποίο θεωρείται ως συνάρτηση του β_1 και ικανοποιεί τις εξισώσεις

$$\left.\frac{\partial Q(\beta)}{\partial \beta_j}\right|_{\beta=\begin{pmatrix}\hat{\beta}_1\\ 0\end{pmatrix}} = 0, \text{ για } j = 1,\ldots,s.$$

Να σημειώσουμε ότι ο $\hat{\beta}_1$ είναι συνεπής εκτιμητής,

$$\left.\frac{\partial L(\beta)}{\partial \beta_j}\right|_{\beta=\begin{pmatrix}\hat{\beta}_1\\ 0\end{pmatrix}} - np'_{a_n}(|\hat{\beta}_j|)\operatorname{sgn}(|\hat{\beta}_j|)$$

$$= \frac{\partial L(\beta_0)}{\partial \beta_j} + \sum_{l=1}^{s}\left\{\frac{\partial^2 L(\beta_0)}{\partial \beta_j \partial \beta_l} + o_p(1)\right\}(\hat{\beta}_l - \beta_{l0})$$

$$-np'_{a_n}(|\beta_{j0}|)\operatorname{sgn}(|\beta_{j0}|) + \{p''_{a_n}(|\beta_{j0}|) + o_p(1)\}(\hat{\beta}_j - \beta_{j0}).$$

Από το Θεώρημα Slutsky καθώς και το Κεντρικό Οριακό Θεώρημα, έχουμε τελικά ότι

$$\sqrt{n}\left(I_1(\beta_{10}) + \Sigma\right)\left\{\hat{\beta}_1 - \beta_{10} + \left(I_1(\beta_{10}) + \Sigma\right)^{-1} b\right\} \to \mathrm{N}\left\{0, I_1(\beta_{10})\right\}.$$

Απόδειξη του Θεωρήματος 2.1.2.1

Πριν προχωρήσουμε στην απόδειξη του θεωρήματος, παρουσιάζουμε δύο λήμματα και τις αποδείξεις τους.

Λήμμα 1. Έστω $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$ να είναι i.i.d. τυχαίες μεταβλητές σύμφωνα με την κατανομή της συνθήκης (A) και μια ακολουθία βαρών $\mathbf{W}_n = (w_{n1}, w_{n2}, \ldots, w_{nn})$ που ικανοποιεί τη συνθήκη (B). Αν $\max\left\{\left|P''_{a_n}(|\ _{j0}|)\right| : \ _{j0} \neq 0\right\} \to 0$ και $\sqrt{n}\neg \to 0$, τότε υπάρχει ένα τοπικό ελάχιστο $\hat{\boldsymbol{\beta}}$ του $Q_n(\boldsymbol{\beta})$ τέτοιο ώστε $\left\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\right\| = O_p(n^{-1/2} + a_n)$, όπου $a_n = \max\left\{\left|P'_{a_n}(|\ _{j0}|)\right| : \ _{j0} \neq 0\right\}$.

Απόδειξη του Λήμματος 1. Έστω $\sim_n = n^{-1/2} + a_n$. Για κάθε $£ > 0$, θέλουμε να δείξουμε ότι υπάρχει μία σταθερά $C > 0$, τ.ώ.

$$P\left\{\inf_{\|\mathbf{u}\| \geq C} Q_n(\boldsymbol{\beta}_0 + \sim_n \mathbf{u}) - Q_n(\boldsymbol{\beta}_0) > 0\right\} \geq 1 - £ .$$

Αυτό συνεπάγεται με πιθανότητα τουλάχιστον $1-£$ ότι υπάρχει ένα τοπικό ελάχιστο τ.ώ. $\left\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\right\| = O_p(a_n)$. Υπενθυμίζουμε ότι η συνάρτηση ποινής SCAD ικανοποιεί $P_a(0) = 0$. Έχουμε ότι:

$$\begin{aligned}
D_n(\mathbf{u}) &= Q_n(\boldsymbol{\beta}_0 + \sim_n \mathbf{u}) - Q_n(\boldsymbol{\beta}_0) \geq L_n(\boldsymbol{\beta}_0 + \sim_n \mathbf{u}) - L_n(\boldsymbol{\beta}_0) \\
&\quad + n\sum_{j=1}^{s}\left\{P_{a_n}(|\ _{j0} + \sim_n u_j|) - P_{a_n}(|\ _{j0}|)\right\} + \frac{n\neg_n}{2}\sum_{j=1}^{p}\left\{(\ _{j0} + \sim_n u_j)^2 - \ _{j0}^2\right\} \\
&= -\sim_n \mathbf{u}^T \mathbf{X}^T \mathbf{W}_n(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}_0) + \frac{1}{2}\sim_n^2 \mathbf{u}^T \mathbf{X}^T \mathbf{W}_n \mathbf{X}\mathbf{u} \\
&\quad + n\sum_{j=1}^{s}\left[\sim_n P'_{a_n}(|\ _{j0}|)\operatorname{sgn}(\ _{j0})u_j + \sim_n^2 P''_{a_n}(|\ _{j0}|)u_j^2\left\{1 + o(1)\right\}\right] \\
&\quad + n\neg_n \sum_{j=1}^{p}\left(\sim_n u_j\ _{j0} + \frac{\sim_n^2}{2}u_j^2\right) = \mathrm{I} + \mathrm{II} + \mathrm{III} + \mathrm{IV}
\end{aligned}$$

Πρώτα δείχνουμε ότι, υπό τις συνθήκες (A) και (B), $\mathrm{I} = O_p(n^{1/2}\sim_n)$. Στην πραγματικότητα, $\mathrm{I} = -\sim_n \mathbf{u}^T \mathbf{X}^T \mathbf{W}_n \boldsymbol{\varepsilon} = -\sim_n \mathbf{u}^T \sum_{i=1}^n w_{ni}\mathbf{x}_i £_i$. Το άθροισμα αποτελείται από ένα διάνυσμα με j στοιχεία ίσα με $\sum_{i=1}^n w_{ni}x_{ij}£_i$. Το $\left\{w_{ni}x_{ij}£_i, i = 1, \ldots, n\right\}$ συγκροτεί μία ακολουθία από ανεξάρτητες μεταβλητές, και κάθε στοιχείο $w_{ni}x_{ij}£_i$ έχει μία κατανομή F_{ni} με μέσο $«_{ni} = E\left(w_{ni}x_{ij}£_i\right) = w_{ni}E\left[x_{ij}E\left(£_i \mid \mathbf{x}_i\right)\right] = 0$

και μία πεπερασμένη διασπορά $\pm_{ni}^2 = E\left(w_i^2 x_{ij}^2 £_i^2\right) = w_i^2 E\left[x_{ij}^2 E\left(£^2 \mid \mathbf{x}_i\right)\right] \le c_2^2 \pm^2 E\left(x_{ij}^2\right) < \infty$. Έστω $B_n^2 = Var\left\{\sum_{i=1}^n w_{ni} x_{ij} £_i\right\}$. Δείχνουμε τώρα ότι η συνθήκη Lindeberg

$$\frac{\sum_{i=1}^n \int_{|t|>¢ B_n} t^2 dF_{ni}(t)}{B_n^2} \to 0, \qquad n \to \infty, \text{ για κάθε } ¢ > 0$$

ικανοποιείται. Στην πραγματικότητα, για κάθε $¢ > 0$, έχουμε

$$\frac{\sum_{i=1}^n \int_{|t|>¢ B_n} t^2 dF_{ni}(t)}{B_n^2} \le \frac{\sum_{i=1}^n w_{ni}^2 \int_{|t/w_{ni}|>¢ c_1 \hat{B}_n / w_{ni}} (t/w_{ni})^2 d\Pr(x_{ij}£_§ \le t/w_{ni})}{c_1^2 \hat{B}_n^2}$$

$$\le \frac{c_2^2}{c_1^2} \times \frac{\sum_{i=1}^n \int_{|t'|>¢' \hat{B}_n} t'^2 dF(t')}{\hat{B}_n^2} \to 0, \qquad \text{όσο } n \to \infty$$

σύμφωνα με τη συνθήκη Lindeberg της i.i.d. τυχαίας ακολουθίας μεταβλητών $\left\{x_{ij}£_i, i = 1, \ldots, n\right\}$, όπου $t' = t / w_{ni}, ¢' = c_1 ¢ / w_{ni}, \hat{B}_n = Var\left\{\sum_{i=1}^n x_{ij}£_i\right\}$ και F είναι η κατανομή των $x_{ij}£_i$. Επομένως, από το κεντρικό οριακό θεώρημα, έχουμε ότι $\mathrm{I} = O_p(n^{1/2}\sim_n)$.

Επίσης, υπό τις συνθήκες (A) και (B), μπορούμε να δείξουμε ότι

$$II = \frac{1}{2}\sim_n^2 \mathbf{u}^T \mathbf{X}^T \mathbf{W}_n \mathbf{X}\mathbf{u} = \frac{1}{2}\sim_n^2 \sum_{i=1}^n w_{ni} \mathbf{u}^T \mathbf{x}_i \mathbf{x}_i^T \mathbf{u} = \frac{1}{2} c_n \sim_n^2 \mathbf{u}^T \sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i^T \mathbf{u}$$
$$= O_p(1) c_n \sim_n^2 n \mathbf{u}^T \operatorname{cov}(\mathbf{X}, \mathbf{X})\mathbf{u} = O_p(n\sim_n^2) \mathbf{u}^T \operatorname{cov}(\mathbf{X}, \mathbf{X})\mathbf{u}, \text{ για } c_1 < c_n < c_2.$$

Επιλέγοντας ένα επαρκώς μεγάλο C, το II μπορεί να επικρατήσει σταθερά πάνω στο I με $\|\mathbf{u}\| = C$. Επιπλέον, από τις συνθήκες του λήμματος, το II επικρατεί εξίσου πάνω στα III και IV. Αυτό ολοκληρώνει την απόδειξη του λήμματος 1.

Αφού, για κάθε ορισμένη τιμή του $\boldsymbol{\beta}_0$, το $a_n = \max\left\{\left|P_{a_n}'(\quad_{j0}\right|\right): \quad_{j0} \ne 0\} = 0$ ικανοποιείται για κάθε επαρκώς μικρό a_n, είναι καθαρό από το λήμμα 1 ότι, επιλέγοντας κατάλληλα a_n και $¬_n$, υπάρχει ένας $\sqrt{n}$ συνεπής ποινικοποιημένος εκτιμητής ελαχίστων τετραγώνων του $\boldsymbol{\beta}$. Η κατάλληλη συνθήκη είναι το $a_n = o(1)$ και το $¬_n$ να είναι ανώτερης τάξης απειροελάχιστο $o(n^{-1/2})$. Στην ανάλυση δεδομένων, αυτό σημαίνει ότι το $¬$ συχνά επιλέγεται να είναι μικρότερο από το a, με n μεγάλο.

Έπειτα θα δείξουμε ότι ο εκτιμητής $\hat{\boldsymbol{\beta}}$ έχει την ιδιότητα της σποραδικότητας.

Λήμμα 2. Έστω $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$ να είναι i.i.d. τυχαίες μεταβλητές σύμφωνα με μια κατανομή που ικανοποιεί τη συνθήκη (A) και μια ακολουθία βαρών $\mathbf{W}_n = (w_{n1}, w_{n2}, \ldots, w_{nn})$ που ικανοποιεί τη συνθήκη (B). Αν

$$\liminf_{a \to \infty} \liminf_{\sim \to 0} P'_{a_n}(\sim) / a_n > 0$$

και για $n \to \infty$

$$a_n \to 0, \qquad \sqrt{n} a_n \to \infty,$$

τότε, για κάθε $\boldsymbol{\beta}$ που ικανοποιεί $\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\| = O(n^{-1/2})$, έχουμε

$$P\left\{ Q_n\left(\boldsymbol{\beta}^{(1)}, \mathbf{0}\right) = \min_{\|\boldsymbol{\beta}^{(2)}\| = O(n^{-1/2})} Q_n\left(\boldsymbol{\beta}^{(1)}, \boldsymbol{\beta}^{(2)}\right) \right\} \to 1.$$

Απόδειξη του Λήμματος 2. Είναι επαρκές να δείξουμε ότι, για κάθε $\boldsymbol{\beta}$ που ικανοποιεί $\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\| = O(n^{-1/2})$ και $j = s+1, \ldots, d$, έχουμε

$$P\left\{ \frac{\partial Q_n(\boldsymbol{\beta})}{\partial \quad_j} < 0, 0 < \quad_j < cn^{-1/2} \right\} \to 1,$$

$$P\left\{ \frac{\partial Q_n(\boldsymbol{\beta})}{\partial \quad_j} > 0, -cn^{-1/2} < \quad_j < 0 \right\} \to 1,$$

και, στην πραγματικότητα,

$$\begin{aligned} \frac{\partial Q_n(\boldsymbol{\beta})}{\partial \quad_j} &= \frac{\partial L_n(\boldsymbol{\beta})}{\partial \quad_j} + \left[nP'_{a_n}(\left| \quad_j \right|) \operatorname{sgn}(\quad_j) + n\lambda \quad_j \right] \\ &= -X_j^T \mathbf{W}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}_0) - X_j^T \mathbf{W}\mathbf{X}(\boldsymbol{\beta} - \boldsymbol{\beta}_0) + nP'_{a_n}(\left| \quad_j \right|) \operatorname{sgn}(\quad_j) + n\lambda \quad_j. \end{aligned}$$

Χρησιμοποιώντας τα επιχειρήματα της ασυμπτωτικής θεωρίας όπως στο Λήμμα 1, βλέπουμε ότι $X_j^T \mathbf{W}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}_0) = X_j^T \mathbf{W}\boldsymbol{\varepsilon} = \sum_{i=1}^n w_{ni} x_{ij} \varepsilon_i = O_p(n^{1/2})$, και το k στοιχείο του $X_j^T \mathbf{W}\mathbf{X}$ είναι $\sum_{i=1}^n w_{ni} x_{ij} x_{ik} \le \sum_{i=1}^n \frac{w_{ni}}{2}(x_{ij}^2 + x_{ik}^2) = O_p(n)$ κάτω από τις συνθήκες (A) και (B). Χρησιμοποιώντας $\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\| = O(n^{-1/2})$, έχουμε

$$\frac{\partial Q_n(\boldsymbol{\beta})}{\partial \ _j} = n^{\mathsf{a}}{}_n \left\{ \mathsf{a}_n^{-1} P'_{\mathsf{a}_n}(\left| \ _j \right|)\mathrm{sgn}(\ _j) + O_p(n^{-\frac{1}{2}} / \mathsf{a}_n) + \frac{\neg_n}{\mathsf{a}_n} \left\| \ _j \right\| \mathrm{sgn}(\ _j) \right\}.$$

Από αυτό, συμπεραίνουμε ότι το πρόσημο της παραγώγου εξαρτάται αποκλειστικά από αυτό του $_j$, και αυτό ολοκληρώνει την απόδειξη του Λήμματος 2.

Οι επιβαλλόμενες συνθήκες του Λήμματος 2 συνεπάγονται ότι κάθε μη σημαντική μεταβλητή των συντελεστών παλινδρόμησης ποινικοποιείται στο μηδέν, δηλαδή $\hat{\boldsymbol{\beta}}^{(2)} = \mathbf{0}$. Αυτή η ιδιότητα εξαρτάται από το a_n και όχι από το $\neg_n$.

Έπειτα, επιστρέφουμε στην απόδειξη του κυρίου θεωρήματος.

Απόδειξη του Θεωρήματος. Για $j = 1, \ldots, s$, έστω

$$\frac{\partial Q_n(\boldsymbol{\beta})}{\partial \ _j}\Big|_{\boldsymbol{\beta}=\begin{pmatrix}\hat{\boldsymbol{\beta}}^{(1)} \\ \mathbf{0}\end{pmatrix}} = 0.$$

Τότε έχουμε

$$\begin{aligned}
0 &= \frac{\partial L_n(\boldsymbol{\beta})}{\partial \ _j}\Big|_{\boldsymbol{\beta}=\begin{pmatrix}\hat{\boldsymbol{\beta}}^{(1)} \\ \mathbf{0}\end{pmatrix}} + nP'_{\mathsf{a}_n}(\left| \hat{\ }_j \right|)\mathrm{sgn}(\hat{\ }_j) + n\neg_n \hat{\ }_j \\
&= -X_j^T \mathbf{W}_n(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}_0) + X_j^T \mathbf{W}_n \mathbf{X} \begin{pmatrix} \hat{\boldsymbol{\beta}}^{(1)} - \boldsymbol{\beta}_0^{(1)} \\ \mathbf{0} \end{pmatrix} + n(P'_{\mathsf{a}_n}(\left| \ _{j0} \right|)\mathrm{sgn}(\ _{j0}) \\
&\quad + P''_{\mathsf{a}_n}(\left| \ _{j0} \right|)(\hat{\ }_j - \ _{j0}) + o_p(1)(\hat{\ }_j - \ _{j0})) + n(\neg_n \ _{j0} + \neg_n(\hat{\ }_j - \ _{j0})) \\
&= -X_j^T \mathbf{W}_n(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}_0) + X_j^T \mathbf{W}_n \mathbf{X}_1(\hat{\boldsymbol{\beta}}^{(1)} - \boldsymbol{\beta}_0^{(1)}) + n\left\{ P'_{\mathsf{a}_n}(\left| \ _{j0} \right|)\mathrm{sgn}(\ _{j0}) + \neg_n \ _{j0} \right\} \\
&\quad + n\left\{ P''_{\mathsf{a}_n}(\left| \ _{j0} \right|) + \neg_n + o_p(1) \right\}(\hat{\ }_j - \ _{j0}).
\end{aligned}$$

Παίρνουμε $\boldsymbol{\Sigma}_n = diag\left\{ P''_{\mathsf{a}_n}(\left| \ _{10} \right|) + \neg_n, \ldots, P''_{\mathsf{a}_n}(\left| \ _{s0} \right|) + \neg_n \right\}^T$ και

$\mathbf{b}_n = \left(P'_{\mathsf{a}_n}(\left| \ _{10} \right|)\mathrm{sgn}(\ _{10}) + \neg_n \ _{10}, \ldots, P'_{\mathsf{a}_n}(\left| \ _{s0} \right|)\mathrm{sgn}(\ _{s0}) + \neg_n \ _{s0} \right)$.

Τότε έχουμε

$$n\left(\mathbf{X}_1^T \mathbf{W}_n \mathbf{X}_1 / n + \boldsymbol{\Sigma}_n + o_p(1) \right)\left(\hat{\boldsymbol{\beta}}^{(1)} - \boldsymbol{\beta}_0^{(1)} \right) + n\mathbf{b}_n = \mathbf{X}_1^T \mathbf{W}_n \left(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}_0 \right).$$

Από το θεώρημα του Slutsky και το κεντρικό οριακό θεώρημα έπεται ότι

$$\sqrt{n}\left(\mathbf{X}_1^T\mathbf{W}_n\mathbf{X}_1/n+\mathbf{\Sigma}_n\right)\left\{\hat{\boldsymbol{\beta}}^{(1)}-\boldsymbol{\beta}_0^{(1)}+\left(\mathbf{X}_1^T\mathbf{W}_n\mathbf{X}_1/n+\mathbf{\Sigma}_n\right)^{-1}\mathbf{b}_n\right\}\to N\left(\mathbf{0},\mathbf{X}_1^T\mathbf{W}_n^2\mathbf{X}_1/n\right).$$

Αυτό ολοκληρώνει την απόδειξη του θεωρήματος.

Απόδειξη της Πρότασης 2.2.2.1

Η ποινή P από τη σχέση 2 μπορεί να γραφτεί ως

$$P_c=\text{a}\left(\boldsymbol{\beta}^T\mathbf{D}^T\mathbf{W}_1\mathbf{D}\boldsymbol{\beta}+\boldsymbol{\beta}^T\mathbf{A}^T\mathbf{W}_2\mathbf{A}\boldsymbol{\beta}\right)$$

όπου $\mathbf{W}_1=diag\left(1/\left(1-\text{°}_{12}\right),1/\left(1-\text{°}_{13}\right),\ldots\right)$ είναι ένας $\left(m\times m\right)$ διαγώνιος πίνακας, με $m=n\left(n-1\right)/2$ να ορίζει τον αριθμό των ζευγαριών $\left(i,j\right),i\neq j$, $\mathbf{W}_2=diag\left(1/\left(1+\text{°}_{12}\right),1/\left(1+\text{°}_{13}\right),\ldots\right)$, ο $\mathbf{D}$ προσδιορίζει τις διαφορές,

$$\mathbf{D}=\begin{pmatrix}1&-1&0&0&\ldots\\1&0&-1&0&\ldots\\\vdots&\vdots&\vdots&\vdots&\ddots\\0&1&-1&0&\ldots\\0&1&0&-1&\ldots\\\vdots&\vdots&\vdots&\vdots&\ddots\end{pmatrix}$$

και $\mathbf{A}$ είναι ο πίνακας που προσδιορίζει την πρόσθεση των παραμέτρων

$$\mathbf{A}=\begin{pmatrix}1&1&0&0&\ldots\\1&0&1&0&\ldots\\\vdots&\vdots&\vdots&\vdots&\ddots\\0&1&1&0&\ldots\\0&1&0&1&\ldots\\\vdots&\vdots&\vdots&\vdots&\ddots\end{pmatrix}.$$

Ο προκύπτων όρος ποινής παίρνει τη μορφή $P_c\left(\boldsymbol{\beta}\right)=\text{a}\boldsymbol{\beta}^T\mathbf{W}\boldsymbol{\beta}$ όπου $\mathbf{W}=\mathbf{D}^T\mathbf{W}_1\mathbf{D}+\mathbf{A}^T\mathbf{W}_2\mathbf{A}$. Μία απλούστερη μορφή του $\mathbf{W}$ λαμβάνεται υπολογίζοντας τις παραγώγους. Έχουμε

$$\frac{\partial P_c\left(\boldsymbol{\beta}\right)}{\partial\ _r}=4\text{a}\sum_{i,r}\frac{1}{1-\text{°}_{ir}^2}\left(\ _r-\text{°}_{ir}\ _i\right)$$

και

$$\frac{\partial P_c(\boldsymbol{\beta})}{\partial \ _r \partial \ _s} = \begin{cases} 4^{\text{a}} \sum_{i \neq s} \frac{1}{1-{}^{\circ 2}_{is}} & \text{αν r=s} \\ -4^{\text{a}} \frac{{}^{\circ}_{rs}}{1-{}^{\circ 2}_{rs}} & \text{αν r} \neq \text{s} \end{cases}$$

που δίνει τον τύπο (4).

Όταν μία συνάρτηση είναι αυστηρά κυρτή, αν ο πίνακας των δεύτερων παραγώγων είναι θετικός, είναι αρκετό να δείξουμε οτι η τετραγωνική μορφή $P_c(\boldsymbol{\beta})$ μηδενίζεται μόνο για $\boldsymbol{\beta} = \mathbf{0}$.

Για ${}^{\circ 2}_{ij} \neq 1, {}^{\text{a}} > 0$, η ποινή $P_c(\boldsymbol{\beta}) / (2^{\text{a}})$ μπορεί να θεωρηθεί ως η τετραγωνική Ευκλείδια νόρμα του επεκταμένου διανύσματος

$$\mathbf{v} = \left(\frac{1^{-}{}_2}{\sqrt{1-{}^{\circ}_{12}}}, \frac{1^{+}{}_2}{\sqrt{1+{}^{\circ}_{12}}}, \frac{1^{-}{}_3}{\sqrt{1-{}^{\circ}_{13}}}, \frac{1^{+}{}_3}{\sqrt{1+{}^{\circ}_{13}}}, \ldots \right).$$

Έτσι, η νόρμα του $\mathbf{v}$ γίνεται μηδέν όταν όλα τα στοιχεία του είναι ίσα με μηδέν. Αυτό πετυχαίνεται μόνο αν ${}_i = 0$ για όλα τα i. Ως εκ τούτου $P_c(\boldsymbol{\beta}) > 0$ αν $\boldsymbol{\beta} \neq 0$. Έτσι η $P_c(\boldsymbol{\beta})$ είναι αυστηρά κυρτή, και το $\hat{\boldsymbol{\beta}}_c$ υπάρχει και είναι μοναδικό.

ΒΙΒΛΙΟΓΡΑΦΙΑ

[1] Aldrich, J. (1997). R. A. Fisher and the Making of Maximum Likelihood 1912-1922, *Statistical Science*, **22**, pp. 162-176.

[2] Anderson, J. A. and Blair, V. (1982). Penalized maximum likelihood estimation in logistic regression and discrimination, *Biometrika,* **69**, pp. 123–136.

[3] Ανδρουλάκης, Ε. (2008). *Μέθοδοι επιλογής μεταβλητών στο μοντέλο αναλογικής διακινδύνευσης του COX και εφαρμογές σε πραγματικά ιατρικά δεδομένα με αποκομμένες παρατηρήσεις.* Μεταπτυχιακή Διπλωματική Εργασία, Εθνικό Μετσόβιο Πολυτεχνείο.

[4] Bickel, P. J. (1975), One-Step Huber Estimates in Linear Models*, Journal of the American Statistical Association*, **70**, pp. 428–433.

[5] Breiman, L. (1998). Arcing classifiers, *Annals Of Statistics,* **26**, pp. 801–849.

[6] Bühlmann, P. (2006). Boosting for high-dimensional linear models, *Annals of Statistics,* **34**, pp. 559–583.

[7] Bühlmann, P. and Yu, B. (2003). Boosting with the L2 loss: Regression and classification, *Journal of the American Statistical Association,* **98**, pp. 324–339.

[8] Donoho, D. L., and Johnstone, I. M. (1994a). Ideal Spatial Adaptation by Wavelet Shrinkage, *Biometrika*, **81**, pp. 425–455.

[9] Donoho, D. L. and Johnstone, I. M. (1995). Adapting to unknown smoothness via wavelet shrinkage, *Journal of the American Statistical Association,* **90**, pp. 1200–1224.

[10] Duffy, D. E. and Santner, T. J. (1989). On the small sample properties of restricted maximum likelihood estimators for logistic regression models, *Communication in Statistics, Theory & Methods,* **18**, pp. 959–989.

[11] Fahrmeir, L. and Kaufmann, H. (1985). Consistency and asymptotic normality of the maximum likelihood estimator in generalized linear models, *The Annals of Statistics,* **13**, pp. 342–368.

[12] Fan, J. (1997). Comments on "Wavelets in statistics a review" by Antoniadis A., *J. Italian Statist. Assoc*., **6**, pp. 131-138.

[13] Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties, *J. Amer. Statist. Assoc*., **96,** pp. 1348-1360.

[14] Frank, I. E., and Friedman, J. H. (1993). A Statistical View of Some Chemometrics Regression Tools, *Technometrics*, **35**, pp. 109–148.

[15] Friedman, J. H. (2001). Greedy function approximation: a gradient boosting ma- chine, *Ann. Statist.,* **29**, pp. 1189–1232.

[16] Friedman, J. H., Hastie, T., and Tibshirani, R. (1999). Additive logistic regression: A statistical view of boosting, *Annals of Statistics,* **28**, pp. 337–407.

[17] Fu, W. J. (1998). Penalized Regression: The Bridge Versus the LASSO, *Journal of Computational and Graphical Statistics*, **7**, pp. 397–416.

[18] Golub, T. R. , Slonim, D. K. , Tamayo, P. , Huard, C. , Gaasenbeek, M. , Mesirov, J. P. , Coller, H., Loh, M. L. , Downing, J. R. , Caligiuri, M. A. , Bloomfield, C. D. , and Lander, E. S. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring, *Science,* **286**, pp. 531–537.

[19] Hastie, T., Tibshirani, R., and Friedman, J. H. (2001). *The elements of statistical learning*, *Springer-Verlag,* New York, USA.

[20]Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: Bias estimation for nonorthogonal problems, *Technometrics,* **12**, pp. 55–67.

[21]Huber P. (1981). *Robust Estimation,* Wiley, New York.

[22]Hurvich, C. M., Simonoff, J. S., and Tsai, C. (1998). Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion, *Journal of the Royal Statistical Society,* B **60**, pp. 271–293.

[23]Klinger, A. (1998). *Hochdimensionale Generalisierte Lineare Modelle*. Ph. D. thesis, LMU München, Shaker Verlag, Aachen.

[24]McCullagh, P., and Nelder, J. A. (1989), *Generalized Linear Models*, 2nd ed., Chapman and Hall, London.

[25]Meir, R. and Rätsch, G. (2003). An introduction to boosting and leveraging. In Mendelson, S. , and Smola, A. (Eds.), *Advanced Lectures on Machine Learning*, pp. 119–184. New York: Springer.

[26]Nelder, J. A. and Wedderburn, R. W. M. (1972). Generalized linear models, *Journal of the Royal Statistical Society,* A **135**, pp. 370–384.

[27]Nyquist, H. (1991). Restricted estimation of generalized linear models, *Applied Statistics,* **40**, pp. 133–141.

[28]Park, M. Y. and Hastie, T. (2007). *An l1 regularization-path algorithm for generalized linear models*, *JRSS*.

[29]Penrose, K. W., Nelson, A. G., and Fisher, A. G. (1985). Generalized body composition prediction equation for men using simple measurement techniques, *Medicine and Science in Sports and Exercise,* **17**, pp. 189.

[30]Schaefer, R. L., Roi, L. D., and Wolfe, R. A. (1984). A ridge logistic estimate, *Communication in Statistics, Theory & Methods,* **13**, pp. 99–113.

[31]Segerstedt, B. (1992). On ordinary ridge regression in generalized linear models, *Communication in Statistics, Theory & Methods,* **21**, pp. 2227–2246.

[32]Shevade, S. K. and Keerthi, S. S. (2003). A simple and efficient algorithm for gene selection using sparse logistic regression, *Bioinformatic,* **19**, pp. 2246–2253.

[33]Siri, W. B. (1956). The gross composition of the body, In Tobias, C. A. & Lawrence, J. H. (Eds.), *Advances in Biological and Medical Physics*, **Volume 4**, pp. 239–280. Academic Press New York.

[34]Tibshirani, R. J. (1996). Regression shrinkage and selection via the LASSO, *J. Roy. Statist. Soc. Ser. B.*, **58**, pp. 267-288.

[35]Tibshirani, R. J. (1997). The Lasso method for variable selection in the Cox model, *Statistics in Medicine*, **16**, pp. 385-395.

[36]Toutenburg, H. (1992). *Lineare Modelle – Theorie und Anwendungen*, Heidelberg: Physica-Verlag.

[37]Trenkler, G. and Toutenburg, H. (1990). Mean squared error matrix comparisons between biased estimators – an overview of recent results, *Statistical Papers,* **31**, pp. 165–179.

[38]Tutz, G. and Binder, H. (2007). *Boosting ridge regression*, *Computational Statistics & Data Analysis*.

[39] Tutz, G. and Leitenstorfer, F. (2007). Generalized smooth monotonic regression in additive modeling, *Journal of Computational and Graphical Statistics,* **16**, pp. 165–188.

[40] Tutz, G. and Ulbricht, J. (2006). *Penalized regression with correlation based penalty,* Discussion Paper 486, SFB 386, Universitat München.

[41] Ulbricht, J. and Tutz, G. (2007). *Boosting correlation based penalization in generalized linear models*, Technical Report Number 009, Department of Statistics, University of Munich.

[42] Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net, *Journal of the Royal Statistical Society,* B **67**, pp. 301–320.